ON MODERN MEASURES AND TESTS OF MULTIVARIATE INDEPENDENCE

Mary Elvi Aspiras-Paler

A Dissertation

Submitted to the Graduate College of Bowling Green
State University in partial fulfillment of
the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2015

Committee:

Maria L. Rizzo, Advisor

Sung Chul Bae,
Graduate Faculty Representative

Junfeng Shang

Wei Ning

ABSTRACT

Maria L. Rizzo, Advisor

For the last ten years, many measures and tests have been proposed for determining the independence of random vectors. This study explores the similarities and differences of some of these new measures and generalizes the properties that are suitable for measuring independence in the bivariate and multivariate case. Some of the measures that brought interest to the statistical community are Distance Correlation (dCor) by Székely and Rizzo [91, 92], Maximal Information Coefficient (MIC) by Reshef et al. [75], Local Gaussian Correlation (LGC) and Global Gaussian Correlation (GGC) by Berentsen and Tjøstheim [6], RV Coefficient by Robert and Escoufier [79] and the HHG test statistic developed by Heller, Heller and Gorfine [42].

This study gives a state-of-the-art comparison of the measures. We compare the measures in terms of their theoretical properties. We consider the properties that are necessary and desirable for measuring dependence such as equitability and rigid motion invariance. We identify which of A. Rényi's postulates [72] can be established or disproved for each measure. Each of the measures satisfies only two if not three properties of Rényi. Among the measures and tests explored in this paper, distance correlation is the only one that has the important characterization of being equal to zero if and only if two random variables or two random vectors are independent.

Several dependence structures including linear, quadratic, cubic, exponential, sinusoid and diamond, are considered. The coefficients of the dependence measures are computed and compared for each structure. The power performance and empirical Type-I error rates of the dependence measures are also shown and compared. For detecting bivariate and multivariate association, dCov and HHG are equally powerful. Both are consistent against all dependence alternatives and the tests achieve good power for finite sample sizes. The RV coefficient is only as powerful as the two previous tests when the relationship is linear.

Dependence measures are applied to real data sets concerning stocks returns and Parkinson's disease.

This is dedicated to...

my husband Ben,

my daughter Bianca,

and most especially,

my father Elpidio Aspiras,

who are my inspiration in everything I do.

ACKNOWLEDGMENTS

I wish to express my heartfelt gratitude to the following people who have helped and supported me in any respect during the completion of this dissertation.

I am sincerely thankful to my advisor, Dr. Maria L. Rizzo, whose guidance and support from the initial to the final stage enabled me to gain understanding of the concepts and helped me attain a great deal of skills. She unselfishly shared her valuable suggestions and insights. She has been my model in Statistics since she became my instructor in Linear Models. She has contributed largely to the study of independence between random vectors which immensely impacted the statistical community. Her expertise on the subject is worthy of great honor.

This thesis would not have been possible without the support of Dr. Wei Ning, Dr. Junfeng Shang and Dr. Sung Chul Bae, who willingly served as my Graduate Committee Members. Their constructive comments help polished the paper.

It is an honor for me to thank the Department of Mathematics and Statistics and the Graduate College for granting me with financial support during my studies at BGSU, without which I will not reach this stage of my career. Specifically, I would like to mention the Graduate Coordinators at the time of my studies: Dr. John Chen and Dr. Hanfeng Chen, who admitted me into the program and provided me with the assistantship; Dr. Rieubert Blok and Dr. Craig Zirbel, who supported until I finished.

I am indebted to many of my professors in the Department of Mathematics and Statistics for sharing their knowledge that has developed my educational outlook. They have been my inspiration to pursue my goals because they have been loyal to their tasks.

It is my pleasure to especially thank Ms. Marcia Lynn Seubert, for all the help in paper works related to graduate students; Ms. Mary Jane Busdeker, for the help in printing and everything that pertains to my student-teaching; and Ms. Barbara Berta, for the assistance with any relevant things.

I am grateful to all my fellow graduate students for their friendship specially Songzi, Doaa, Junvie, Olu, Grace, who have helped and supported me in one way or another. I want to specially thank Ying-Ju for her unselfish kindness, and help in whatever I ask of her, even in editing my

R-code. She is such a blessing. I would like to mention my special friends Si Shi, Ramadha, Kevin and all my classmates I fail to mention, who academically grow with me as I pursue this education.

I would like to extend my gratitude to my friends outside of BGSU and lifegroup specially Loni and Tom, Tobi and Funmi, Alma, Thea, Gela, Joy, Jonah, Kayode, Katie, Calvin, Christian, Efua, Morayo, Itunu, Esther, Eva, Naa, Yaa, Lily, Fang, Quijan, Nikki, Evans, Raphael, Manuel, Summer, Tobi, Tolu and many others, for always being there to pray and support. Also, special thanks to my Filipino community including Tita Elsie, Tita Lily and Tito Rudy, Tito Mario and Tita Josie, Tito Roger and Tita Linda, Tito Alex and Tita Babie, Tito Ramon and Tita Emily, Tito Noli and Tita Josie, Francis, Jorge and Cristina and many others, who supported me in many, many ways.

I would like to give my appreciation to my family (Papa, Tita, Emily, Natasha and Nathalie) and my husband's family (Papa Jun, Mama Vebie, Juby, Gerly, Beth and Ron) for believing in me and for supporting me through prayers and encouragement.

I am so thankful for the love and support of my husband, Ben and my precious daughter, Bianca who inspired me to pursue and finish. Their presence provided me the strength to continue to labor.

Most of all, I honor the One who grant me knowledge and wisdom.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 INTRODUCTION

Knowledge of the association or relationship between two random variables or two random vectors is very important in the statistical world. There have been many measures and tests that were developed over the years to provide tools to accurately identify whether certain relationships occur. In the past decade, several new multivariate methods have appeared in the literature. Most of these measures and tests possess many different properties that are useful for various types of data. Hence, this study aims to examine modern measures and tests that are available.

The focus of this research is to discuss and compare five recently developed measures and tests of independence that brought interest to the statistical community. These are distance correlation by Székely and Rizzo [91, 92], Global Gaussian correlation by Berentsen and Tjøstheim [6], the maximal information coefficient by Reshef, Reshef, Finucane, Grossman, McVean, Turnbaugh, Lander, Mitzenmacher and Sabeti [75], the RV coefficient by Robert and Escoufier [79], and the HHG test statistic developed by Heller, Heller and Gorfine [42]. We identify their advantages and disadvantages and examine the properties of each. We include the definition and properties of the classical measure of correlation which is the Pearson product-moment correlation as well as the maximal correlation.

It has been widely known that the Pearson correlation coefficient is a bivariate measure of linear association of two random variables. However, it fails to capture nonlinear dependence structures in bivariate data, hence its accompanying test tends to have low power in these cases. Though it works well for approximately bivariate Gaussian variables, a drawback of this coefficient is that for non-Gaussian random variable, the population correlation ρ can be equal to 0 even when the variables are dependent. A new method constructed by Tjøstheim and Hufthammer [99] around the concept of Gaussian correlation is the global Gaussian correlation. This coefficient is derived from a local correlation function which is based on approximating a bivariate density locally by a family of bivariate Gaussian densities utilizing local likelihood. That is, at each point of the distribution, a Gaussian distribution is approximated and the correlation of the approximating Gaussian

distribution is taken as the local correlation in that neighbourhood. This procedure gives a precise mathematical description and interpretation of correlation problems particularly in finance and economics where association between financial objects become stronger as the market declines in behavior with correlation approaching one when the market collapses. A global measure of dependence was then developed by Berentsen and Tjöstheim [6] by aggregating the local correlations $\rho(x, y)$ on subsets of $\mathbb{R}^2$.

Another interesting coefficient measure that will be considered is the normalized coefficient of distance covariance (dCov) called distance correlation (dCor). This empirical distance dependence measure is based on functions of Euclidean distances between sample elements rather than sample moments. It measures dependence between random vectors X and Y in arbitrary dimension, can detect nonlinear complex dependence, and the corresponding dCov test is consistent for any types of alternatives. Thus, the test has power against non-monotone relationships. They are analogous to, but more general than Pearson product-moment covariance and correlation. A fundamental property of distance correlation is that its value being zero characterizes independence of X and Y.

The maximal information coefficient (MIC) is built for data exploration. It is a binning method which is based on mutual information values. It identifies a subset of strongest associations in a large data set that contains too many closely related pairwise associations that are hard to delve into manually. It belongs to a larger class referred to as MINE (Maximal Information-based Nonparametric Exploration) statistics [75], which have been proposed for identifying and classifying relationships. Reshef et al. claimed this method to be highly equitable. However, there have been numerous questions on the claim of MIC having the property of equitability. The definition of equitability can be found in Section 4.2.1. Simon and Tibshirani [85] showed in their simulations that MIC has serious power deficiencies and that it will produce too many false positives when it is used for large-scale exploratory analysis. Moreover, if it has low power, they concluded that the equitability property of MIC is not very useful.

In addition, we included the RV coefficient because it is a multivariate generalization of the

squared Pearson correlation coefficient that takes values between 0 and 1. It makes use of the ideas of variance and covariance of vector-valued random variables, where it measures the proximity of two sets of points that may each be represented in a matrix. The rationale of RV coefficient is to consider that two sets of variables are correlated if the relative position of the samples in one set is similar to the relative position of the samples in another set. Josse and Holmes [51] wrote a review about the RV coefficient as well as the distance correlation. The authors [50] discussed ways to test the significance of the RV coefficient. It has been used in many fields such as sensory analyses, morphology, and neuroscience and it has been proposed by Robert and Escoufier [79] as a unifying tool for linear multivariate statistical methods including principal component analysis, discriminant analysis and correlation analysis.

The last test to be assessed is not a measure but a multivariate test of association based on ranks of distances developed by Heller, et al. [42]. It is used to detect associations between random vectors of any dimensions based on pairwise distances of the values of X and Y. The statistic is a function of ranks of these distances.

Each of these coefficients is discussed thoroughly in Chapter 3.

Seven desirable properties of dependence measures outlined by A. Rényi [72] found in Chapter 4 are verified for these recent developed measures. Rényi verified these properties on some known measures like correlation coefficient, correlation ratios, the mean square contingency and the maximal correlation. He showed that while the first three measures satisfy some properties, only the maximal correlation satisfies all seven of them. Two years before Rényi established these properties, Linfoot [63], constructed a measure of dependence based on the amount of information $I(X, Y)$ which X and Y contain with respect to each other. The quantity is given by $L(X, Y) = (1 - \exp(-2I(X, Y)))^{\frac{1}{2}}$. This quantity satisfies all seven of the properties that Rényi proposed [72]. Some authors such as Schweizer and Wolff [81] and Granger, Maasoumi and Racine [38] partially modified these properties to suit the statistics they establish.

Also in the same chapter, other desirable properties of these measures are assessed. More specifically, rigid motion invariance and equitability property were investigated. Rigid motion in-

variance means that the dependence measure is invariant to distance-preserving transformations such as translations, rotations and reflections. Equitability, on the other hand, is a property that enables the dependence measure to give similar scores to equally noisy relationships regardless of relationship types. Authors of MIC proposed that a good measure of dependence should be equitable just like MIC; however, they have not proven the equitability property theoretically, but only analyzed it using the results of simulated data. Kinney and Atwal [54] showed that the authors' simulation evidence is artifactual and offered mathematical proof that neither MIC nor any non-trivial dependence measure satisfies the definition of equitability. However, the heuristic notion of equitability can instead be formalized using a self-consistent and more general definition that follows naturally from the Data Processing Inequality. This concept will be discussed further in Chapter 4.

In Chapter 5, several dependence structures such as linear, quadratic, cubic, exponential and sinusoid are considered and simulated and the performance of these measures of dependence are compared. Power comparisons and Type I error rates are presented for each case.

Then in Chapter 6, the dependence measures were applied and compared on real data sets of stock market analysis and Parkinson's disease.

Finally, summary and conclusions are provided in Chapter 7.

CHAPTER 2 LITERATURE REVIEW

2.1 Origin of Correlation

There are numerous papers, articles and books about the study of dependence. Many measures and test statistics have been formulated and they have evolved over the years. The concepts and methods are substantial, some are simple, others are complex, but generally evaluation of these measures leads to a common set of criteria. We have observed that these criteria usually include accuracy and computational simplicity. In this chapter, we review past measures and tests with their properties.

The word "correlation" first appeared when Galton [28] presented a paper to the Royal Society on December 5, 1888, entitled "Correlations and their Measurement chiefly from Anthropometric Data." Galton read the opening lines as follows: "Co-relation or correlation of structure is a phrase much used in biology, and not least in that branch of it which refers to heredity, and the idea is even more frequent than the phrase; but I am not aware of any previous attempt to define it clearly, to trace its mode of action, or to show how to-measure its degree." He defines correlation in the same paper, "Two variable organs are said to be correlated when the variation of the one is accompanied on the average by more or less variation of the other, and in the same direction. It is easy to see that co-relation must be the consequence of the variations of the two organs being partly due to common causes. If they were in no respect due to common causes, the co-relation would be nil." Though Galton did not introduce the idea of negative correlation, he revealed the properties of the correlation coefficient. Also, his method was still imprecise by modern standards but it was embraced by other researchers. It was later that Karl Pearson [67] provided the familiar mathematical framework of correlation.

Since then, correlation brought interest to the mathematicians, statisticians and psychologists who use this measure.

In 1904, Spearman [88] attempted to remedy the deficiency of scientific correlation to cater to

the need of the practical workers. He mentioned that a good method of correlation should have the following requirements:

1. Quantitative expression - the most fundamental requisite is to be able to measure the observed correspondence by plain numerical symbol.

2. Significance of the quantity - this means that a measure might be afforded of the hidden underlying cause of the variations.

3. Accuracy - should be truly representative of the sample. In measuring the correlation, one must carefully distinguish the variety of ways of calculating any one of them. The smallness of probable error depends principally upon the number of cases observed but also largely upon the mathematical method of correlation. The best method is that one which gives the least probable error other things being equal.

4. Ease of application - In addition to a standard method, which must be used for finally establishing principal results, there is an urgent need also of auxiliary methods capable of being employed under the most varied conditions and with the utmost facility.

The requirements being mentioned imply that Spearman already had in mind the criteria of an alternate measure of association. He stated in the paper the advantages and disadvantages of using the "rank" method. He formulated the Spearman's rank correlation coefficient or Spearman's rho. It is based on the ranks of the n raw scores of observations X_i and Y_i. It is defined as the Pearson correlation coefficient between the ranked variables.

2.2 Progression of Correlation

Wilks [101] considered a criterion for testing the mutual independence of k sets of normally distributed variables. The criterion is a function of observations only and is derived as a Neyman-Pearson λ ratio by applying the principle of maximum likelihood. It is expressed as the ratio of the determinant of the matrix of correlation coefficients of all variables to the product of the

determinants of the correlation coefficients within the k sets. This ratio is used to study the relationships among the variables separately when the observations on several variables are taken into account simultaneously. The relationships of these variables are to be taken into account; otherwise a considerable part of the information supplied by the observations are lost. Until now, Wilks' lambda has a wide range of application, as it is used in multivariate hypothesis testing such as a likelihood-ratio test and multivariate analysis of variance (MANOVA). The likelihood ratio test is not applicable if the dimension exceeds the sample size or when the distributional assumption does not hold.

After a year, Hotelling and Pabst [46] defined a rank correlation coefficient of two continuous variates that are not likely to be normally distributed. They found the most convenient formula for computing the coefficient, which appears to be a sensitive index of relationship, since for a given value of n, it possesses a greater number of distinct values. This coefficient is regarded chiefly as a more easily calculable substitute for the product-moment correlation coefficient r. However, Pearson has remarked that the rank correlation coefficient is easier to compute for samples that are smaller than approximately forty (40), while product-moment correlation r is more convenient for larger samples. We can see here that the developers are not only concerned with concept and theory but the computability of a statistic.

It was then succeeded by Kendall [53] when he developed a rank correlation measure, τ, that is based on the actual score of any given ranking and the maximum possible score of the observations, when they are all in the objective/correct order. It produced a ratio and there are two methods for the calculations of τ. It is easy to calculate. It is a logical measure of ranking carried out by the process and proves useful in psychological work.

Another test that is based only on the rank order of the observations was devised by Hoeffding [45]. It makes use of a random sample of size n with continuous distribution function. The test is consistent with respect to the class Ω of distribution functions with continuous joint and marginal probability densities.

Blum, Kiefer, and Rosenblatt [11] are famous to statisticians that utilize empirical distribution

functions in testing independence. They discussed certain tests of independence based on the sample distribution function that possess power properties superior to other tests of independence. These tests, developed in view of the Cramér-von Mises tests, are based on large values of

$$B_n = \int (T_n(r))^2 dS_n(r)$$

where $T_n = S_n(r) - \prod_{j=1}^{m} S_{n_j}(r_j)$, S_n is the sample distribution function of independent random m-vectors $X_1, \cdots, X_n$ with common unknown distribution function F, and S_{n_j} is the marginal distribution function associated with the j^{th} component of the X_i. The characteristic functions of the limiting distribution functions of a class of such test criteria are obtained and the corresponding distribution function is tabled in the bivariate case $(m = 2)$. The tests have asymptotic normal distributions and when $m = 2$, it is equivalent to the test proposed by Hoeffding [45].

Blomqvist [10] realized that it frequently happens that it is unnecessary to utilize all the information given by the data. In such cases, he said that it is desirable to use methods which are:

1. valid under rather weak assumptions regarding the distribution of the population

2. easy to deal with in practice.

However, in most cases, the applicability of such methods is limited by their small efficiency. Some known rank correlation coefficients such as Hoeffding [45], Kendall [53], or Spearman [88] only satisfy the first property, and in large sample sizes, these coefficients are not easy to calculate. Therefore, Blomqvist formed a simpler method called q' for testing independence. The q' statistic is based on the number of sample points n_1 and n_2 that belongs to the first and third quadrants, respectively, where the x, y-plane is divided into four regions by the lines $x = \bar{x}$ and $y = \bar{y}$. It is asymptotically normally distributed. The coefficient q' has both properties mentioned above and can be used whenever its efficiency is not too small. A test based on q' is nonparametric and its asymptotic efficiency in the normal case is about 41 percent. It is similar to the special case of the exact test of independence in the 2×2 table by Fisher [26].

Kruskal [56] also contributed to the study of ordinally invariant measures of association for bivariate populations. But he focused on the probabilistic and operational interpretations of the population values. He discussed the relationships and connections of the quadrant measure of Blomqvist, Kendall's τ and Spearman's ρ and certain other measures of association for cross classifications.

Bhuchongkul [7] continued to study the class of rank tests for independence. He developed a test that is based on the following form:

$$T_n = \frac{1}{N}\sum_{i=1}^{N} E_{N,r_i} E'_{N,s_i} Z_{N,r_i} Z'_{N,s_i}$$

where E_{N,r_i}, E'_{N,s_i}, $i = 1, \cdots, N$ are two sets of constant satisfying certain restrictions and $Z_{N,r_i} = 1(Z'_{N,s_i} = 1)$ when $X_i(Y_i)$ is the $r_i th(s_i th)$ smallest of the $X's(Y's)$ and $Z_{N,r_i} = 0(Z'_{N,s_i} = 0)$ otherwise. When E_{N,r_i}, E'_{N,s_i} are taken as the expected value of the $r_i th(s_i th)$ standard normal order statistic from a sample size of N, they got the normal scores test statistic. If they replace $E_{N,r_i} = r_i$ and $E'_{N,s_i} = s_i$, the resulting test statistic is equivalent to the Spearman rank correlation statistic. The normal scores test is shown to be (a) the locally most powerful rank test and (b) asymptotically as efficient as the parametric correlation coefficient for some specified alternatives when the underlying distributions are normal. Sinha and Wieand [86] extended Bhuchongkul's bivariate rank statistics into multivariate generalizations for testing multivariate independence. It is shown that the test statistics can be expressed a a rank statistics which are easy to compute, have asymptotic normal distributions and can detect mutual dependence in alternatives which are pairwise independent.

Another statistic by Geiser and Randles [31] is $\hat{Q}_n$. It is based on interdirections used for testing whether two vector-valued quantities are dependent. Counts, called interdirections, measure the angular distance between two observation vectors relative to the positions of the other observations. The statistic $\hat{Q}_n$ has an intuitive invariance property. It is resistant to outliers. It has a limiting chi-squared distribution under the null hypothesis of independence when each vector is elliptically

symmetric. It compares favorably to Wilks' likelihood ratio criterion when the vectors are heavy-tailed elliptically symmetric distributions. It performs uniformly better than the componentwise quadrant statistic of Blomqvist [10] when the vectors are spherically symmetric. It is better than the others for heavy-tailed distributions and is competitive for distributions with moderate tail weights. It reduces to the quadrant statistic when the two quantities are each univariate. It also reduces to the sample coefficient of medial correlation in the bivariate case and so is a natural extension of a simple sign statistic.

An extension of the quadrant test statistic of Blomqvist [10] based on spatial signs is proposed by Taskinen, Kankainen and Oja [97] for testing independence. It has the property of affine invariance. It is asymptotically equivalent to the interdirection test by Geiser and Randles [31] in the elliptic case. Moreover, Taskinen, Oja and Randles [98] proposed test statistics for testing independence between two random vectors. This is a sequel to the multivariate extension work of Gieser and Randles as well as Taskinen, Kankainen and Oja. Multivariate extensions of Kendall's tau and spearman's rho statistics are presented using two different approaches. First, interdirection proportions are used to estimate the cosines of angles between centered observations and between differences of observation vectors. Second, covariances between affine-equivariant multivariate signs and ranks are used. Both test statistics produced appear to be asymptotically equivalent if each vector is elliptically symmetric. The spatial sign versions are easy to compute for data in common dimensions and they provide practical, robust alternatives to normal theory methods. Simple algorithms were formulated for easy computation.

Moreover, there are measures that are robust against non-normality. These are Spearman-rank correlation, Kendall-tau correlation, Fisher-Yates normal scores and nonparametric curve estimation techniques. These tests, which are designed mostly for bivariate dependence, are not consistent. Puri and Sen [68] developed a coefficient but it is not applicable for random variables with dimensions greater than the sample size. Taskinen, Oja, and Randles [98] came up with a measure for higher dimensions that are based on component-wise ranking but are ineffective for testing non-monotone types of dependence. Some authors have also found graphical presentations that

show association. Correlographs by Feuerverger [23], Corrgrams by Friendly [27], Dependogram by Bilodeau and Lafaye de Micheaux [9], and Dependence Maps by Jones and Koch [49] are a few of them.

2.3 Recent Development of Correlation

A unique nonparametric approach to the problem of testing the joint independence of two or more random vectors is developed by Bakirov, Rizzo, and Székely [3]. Distance covariance is based on a measure of association determined by interpoint distances. It does not require distributional assumptions or continuity, and does not require computing the inverse of the covariance matrix. Distance correlation has nice properties such as (a) the population independence coefficient takes values between 0 and 1, (b) the coefficient equals zero if and only if the vectors are independent. The corresponding statistic has a finite limit distribution if and only if the two random vectors are independent and diverges to infinity stochastically as sample size $n \to \infty$. Hence, a universally consistent test is determined by the statistic. It is applicable in arbitrarily high dimension and very powerful against non-monotone types of dependence. The exact distance correlation coefficient in the bivariate normal case is an increasing function of the absolute value of the product moment correlation and coincides with the absolute value of correlation in the Bernoulli case. A modification of the statistic makes it affine invariant. The independence coefficient and the proposed statistic both have a natural extension to testing the independence of several random vectors.

Several tests of independence are derived from the empirical characteristic function. In addition to the one developed by Csörgő [18] and Bakirov et al. [3], is a consistent bivariate nonparametric test of dependence by Feuerverger [23]. Feuerverger's proposed test is based on the same L^2 distance function as distance covariance (dCov), but applies ranks. Difficulties were noted for this particular test in extending to higher dimensions. Another one is proposed by Bilodeau and Lafaye de Micheaux [8]. The test statistic is a Cramér-von Mises functional of a process defined from the empirical characteristic function. It is used to test independence between marginal vectors each of which is normally distributed but without assuming the joint normality of these marginal vectors.

It can be represented as a V-statistic and it is said to be consistent to detect any form of dependence.

Copula-based measures of dependence were first exploited by Schweizer and Wolff [81] when Sklar [87] introduced the notion of copula. They defined several nonparametric measures of dependence for pairs of random variables using copulas. They showed that these measures satisfy some modifications of the set of axioms for measures of dependence proposed by Rényi. They showed that the copula of a pair of random variables X, Y is invariant under almost surely strictly increasing transformations of X and Y and that any property of the joint distribution function of X and Y which is invariant under such transformations is solely a function of their copula. They made slight modifications of Rényi's axioms $\mathbf{E}$, $\mathbf{F}$ and $\mathbf{G}$. For example, in axiom $\mathbf{E}$, they restrict their components to pairs of continuously distributed random variables replacing "if" by "if and only if" and limiting f and g to be almost surely strictly monotone functions. For axiom $\mathbf{G}$, they allow $\delta(X, Y)$ to be a strictly increasing function of the absolute value of Pearson's $r(X, Y)$. They also added axiom $\mathbf{H}$ for continuity property.

Siburg and Stoimenov [84] also constructed via the distance between the copula of continuous X and Y and the independent copula. They called this a measure of mutual complete dependence (m.c.d.). They defined that on a common probability space, two random variables X and Y are mutually completely dependent if each variable is a function of the other with probability one. The measure they constructed is a natural approach that crucially depends on the choice of the distance function. They used a modified Sobolev norm, with respect to which mutual complete dependence cannot approximate any other kind of dependence. The Sobolev norm produces the first nonparametric dependence which precisely captures the two extremes of dependence:

1. It is 0 if and only if X and Y are independent.

2. It is 1 if and only if X and Y are m.c.d.

Another well-known measure of statistical dependence between two random variables is the mutual information. For two random variables X and Y with joint probability distribution, the definition of mutual information was given by Shannon and Weaver [83] and Cover and Thomas

[17]. It is related to the Kullback-Leibler divergence between the joint density and the product of the marginal densities. The mutual information is zero if and only if the random variables are independent. It follows that the mutual information captures all dependencies between random variables not just say, second order ones as captured by the covariance. It is symmetric, and additive for independent variables.

Granger, Maasoumi and Racine [38], who considered a robust nonparametric implementation of a metric entropy measure of association and dependence, mentioned that a good measure of dependence should satisfy the following properties:

1. It is well-defined for both continuous and discrete variables.

2. It is normalized to zero if X and Y are independent, and lies between 0 and 1.

3. The modulus of the measure is equal to unity (or a maximum) if there is a measurable exact (nonlinear) relationship, $Y = m(X)$ say, between the random variables.

4. It is equal to or has a simple relationship with the (linear) correlation coefficient in the case of a bivariate normal distribution.

5. It is metric, that is, it is a true measure of distance not just of divergence.

6. The measure is invariant under continuous and strictly increasing transformations $\varphi(\cdot)$. This is important since X and Y are independent if and only if $\varphi(X)$ and $\varphi(Y)$ are independent.

All the above properties encompass the advantages of their method. Notice that some properties particularly numbers 2, 3 and 4, are similar to Rényi's postulates.

A dependence measure expressed in terms of principal components are proposed by Delicado and Smrekar [19] to measure not necessarily linear related variables. They used the covariance and linear correlation as measures of local linear relationship and generalize these for variables distributed along a curve. They determined which properties of Rényi are satisfied by their new measures. All, except **F**, with some modifications of **A**, **D** and **E** were verified. For instance, they modified axiom **A** by saying that $\delta(X, Y)$ is defined for any pair of random variables X and

Y distributed along a curve according to the their definition with generating variables (S, T) such that the product of the local variances of S and T which is given by $LV_X(S)LV_Y(T)$ is not equal to 0 with probability 1.

A most recent innovation that depends on the nature of function spaces is the reproducing kernel Hilbert spaces (RKHSs). Gretton, Herbrich, Smola, Bousquet and Schölkopf [39] introduce new two kernel-based independence measures, the constrained covariance (COCO) and the kernel mutual information (KMI), to measure the degree of independence of random variables. COCO is defined simply as the spectral norm of the covariance operator between RKHSs while KMI, a more sophisticated measure, is a function of the entire spectrum of the covariance operator. These two quantities are both based on the covariance between functions of the random variables in RKHSs. They proved that when the RKHSs are universal, COCO and KMI is 0 if and only if the random variables being tested are independent. They also showed that the KMI is an upper bound near independence on the Parzen window estimate of the mutual information, which becomes tight (i.e. zero) when the random variables are independent. It was shown also that same results apply for two-correlation based dependence functionals. That is, the kernel canonical correlation (KCC) and the kernel generalised variance (KGV) are independence measures for universal kernels and proved the latter to be an upper bound on the mutual information near independence.

Furthermore, Gretton, Bousquet, Smola and Schölkopf [40] proposed an independence criterion based on the eigen-spectrum of covariance operators in RKHSs, consisting of an empirical estimate of the Hilbert-Schmidt norm of the cross-covariance operator. They called it Hilbert-Schmidt Independence Criterion (HSIC). Compared to the previous kernel-based independence criteria Gretton et al. developed above, this criterion has several advantages. First, the empirical estimate is simpler than any other kernel dependence test and requires no user-defined regularisation or tuning beyond kernel selection. Second, the empirical estimate converges to the population quantity at the rate $1/\sqrt{m}$ where m is the sample size. In particular, it is guaranteed to detect any existing dependence with high probability. Third, HSIC is much more robust to outliers. Finally, in their empirical studies, HSIC has performance that met or exceeded the methods compared on

all data sets besides the case of $m = 250$.

Reimherr and Nicolae [71] presented a framework for selecting and developing measures of dependence when the goal is the quantification of a relationship between two variables, not simply the establishment of its existence. For them, the most significant creative applications for dependence measures are

1. detection - detecting dependence of any form
2. ranking - ordering the dependence in different relationships
3. quantification - summarizing a relationship in an informative fashion.

In contrast to the six axioms of a good measure enumerated by Rényi [72], they proposed only three guidelines for quantifying dependence:

1. existence - the measures should exist for a large collection of random variables, vectors and/or functions
2. range - the range of measure should be $[0, 1]$
3. interpretability - the measure should have a clear interpretation, for all possible values, based on information content.

Since interpretability is the most crucial part, they incorporated the interpretability property into a measure they propose. They introduced a function they called *information link function* that measures the amount of information as determined by the practitioner and the problem involved. Hence, it is the responsibility of the researcher to determine if a given information function has an interpretation relevant to their analysis. In short, they demonstrated how more general measures of information can be used to achieve the same goal. They devised a plan on how to build dependence measures that is designed to allow practitioners to tailor measures to their needs. They showed three examples that fall naturally into their framework. The first two are common in statistics and these are reflecting prediction and statistical efficiency. The third one which is related to information theory is entropy.

As what we have seen, there are many measures and tests that can be helpful and productive in analyzing many data types. We agree with what Gelman said in his blog post [34], that he imagines that these different measures of dependence could be useful for different purposes.

CHAPTER 3 METHODS AND PROPERTIES OF DEPENDENCE MEASURES

3.1 Distance Correlation

3.1.1 Distance Covariance and Distance Correlation

Distance correlation ($\mathcal{R}$ or dCor) provides a new approach to the problem of measuring dependence and testing the joint independence of two random vectors $\mathbf{X}\epsilon\mathbb{R}^p$ and $\mathbf{Y}\epsilon\mathbb{R}^q$. It is a new measure of dependence developed by Székely, Rizzo and Bakirov [91]. They showed that for all random variables with finite first moments, the distance correlation generalizes the idea of correlation in three ways:

1. $\mathcal{R}$ is defined for two random vectors X and Y of arbitrary dimensions, not necessarily equal.

2. $\mathcal{R} = 0$ if and only if the random vectors are independent.

3. It does not require distributional assumptions.

The distance covariance and distance correlation are analogous to product-moment covariance and correlation, respectively, but give a more general idea as measures of multivariate independence. These empirical distance dependence measures are based on functions of Euclidean distances between sample elements rather than sample moments. The distance covariance $\mathcal{V}$ can be applied to measure the distance $|f_{X,Y}(t,s) - f_X(t)f_Y(s)|$ between the joint characteristic function of X and Y denoted by $f_{X,Y}$ and the product of the marginal characteristic functions denoted by f_X and f_Y. It is a weighted L_2 measure in which the choice of suitable weight function is crucial to ensure the property of independence. The weight function which the authors used here is a special one since this is the only weight function (among the positive weight functions for which the L_2 norm exists) such that the weighted L_2 distance defined below is rigid motion invariant and scale equivariant.

Definition 3.1. *The distance covariance (dCov) is defined as*

$$\mathcal{V}^2(X,Y;w) = \int_{\mathbb{R}^{p+q}} |f_{X,Y}(t,s) - f_X(t)f_Y(s)|^2 w(t,s)dtds \tag{3.1.1}$$

where $w(t,s) = (c_p c_q |t|_p^{1+p} |s|_q^{1+q})^{-1}$ *and* $c_k = \frac{\pi^{(1+k)/2}}{\Gamma((1+k)/2)}$ *and* $\Gamma(\cdot)$ *is the complete gamma function.*

This definition is analogous to the classical covariance but it has a special property that $\mathcal{V}^2(X,Y;w) = 0$ if and only if X and Y are independent.

A standard version of $\mathcal{V}(X,Y)$ is the distance correlation defined as follows.

Definition 3.2. *The distance correlation (dCor) between random vectors* X *and* Y *with finite first moments is the nonnegative number* $\mathcal{R}(X,Y)$ *defined by*

$$\mathcal{R}^2(X,Y) = \begin{cases} \frac{\mathcal{V}^2(X,Y)}{\sqrt{\mathcal{V}^2(X)\mathcal{V}^2(Y)}}, & \mathcal{V}^2(X)\mathcal{V}^2(Y) > 0; \\ 0, & \mathcal{V}^2(X)\mathcal{V}^2(Y) = 0. \end{cases} \tag{3.1.2}$$

where $\mathcal{V}^2(X)$ *and* $\mathcal{V}^2(Y)$ *are the distance variance of* X *and distance variance of* Y*, respectively.*

The distance correlation is rigid motion invariant [93] because distances are invariant to location and scale transformations.

The distance variance of X and distance variance of Y are each defined similarly as the distance covariance given by

$$\mathcal{V}^2(X;w) = \int_{\mathbb{R}^{2p}} |f_{X,X}(t,s) - f_X(t)f_X(s)|^2 w(t,s)dtds, \tag{3.1.3}$$

$$\mathcal{V}^2(Y;w) = \int_{\mathbb{R}^{2q}} |f_{Y,Y}(t,s) - f_Y(t)f_Y(s)|^2 w(t,s)dtds. \tag{3.1.4}$$

Definition 3.3. *The empirical distance covariance* $\mathcal{V}_n(\mathbf{X},\mathbf{Y})$ *is the nonnegative square root of*

$$\mathcal{V}_n^2(\mathbf{X},\mathbf{Y}) = \frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}, \tag{3.1.5}$$

where $A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$, $a_{kl} = |X_k - X_l|_p$, $\bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^{n} a_{kl}$, $\bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^{n} a_{kl}$ *and* $\bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,l=1}^{n} a_{kl}$.

Similarly, the empirical distance variances $\mathcal{V}_n(\mathbf{X})$ and $\mathcal{V}_n(\mathbf{Y})$ are defined as the nonnegative square roots of

$$\mathcal{V}_n^2(\mathbf{X}) = \frac{1}{n^2}\sum_{k,l=1}^{n^2} A_{kl}^2, \tag{3.1.6}$$

$$\mathcal{V}_n^2(\mathbf{Y}) = \frac{1}{n^2}\sum_{k,l=1}^{n^2} B_{kl}^2 \tag{3.1.7}$$

Definition 3.4. *The empirical distance correlation* $\mathcal{R}_n(\mathbf{X}, \mathbf{Y})$ *is the square root of*

$$\mathcal{R}_n^2(\mathbf{X}, \mathbf{Y}) = \begin{cases} \frac{\mathcal{V}_n^2(\mathbf{X},\mathbf{Y})}{\sqrt{\mathcal{V}_n^2(\mathbf{X})\mathcal{V}_n^2(\mathbf{Y})}}, & \mathcal{V}_n^2(\mathbf{X})\mathcal{V}_n^2(\mathbf{Y}) > 0; \\ 0, & \mathcal{V}_n^2(\mathbf{X})\mathcal{V}_n^2(\mathbf{Y}) = 0. \end{cases} \tag{3.1.8}$$

Clearly, $\mathcal{R}_n(\mathbf{X}, \mathbf{Y})$ is computationally simple. According to Newton [66], it satisfies Don Geman's *elevator test* because the method can be explained to a colleague at the same time it takes an elevator to go between floors.

An equivalent definition of dCov, according to Székely and Rizzo [92], is that if $E|X| < \infty$ and $E|Y| < \infty$, then

$$\mathcal{V}^2(X, Y) = E(|X - X'||Y - Y'|) + E(|X - X'|)E(|Y - Y'|) - 2E(|X - X'||Y - Y''|).$$

It can be shown that

$$\mathcal{V}^2(X, Y) = cov(|X - X'|, |Y - Y'|) - 2cov(|X - X'|, |Y - Y''|)$$

where X', Y' and Y'' are independent copies of X and Y whereas $|X - X'|$ and $|Y - Y'|$ are Euclidean distances.

The empirical value of distance covariance is equivalent to the L_2 norm of the difference of the empirical characteristic function of the sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ and the marginal empirical characteristic functions of the sample X and the sample Y. That is,

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = |f_{X,Y}^n(t, s) - f_X^n(t) f_Y^n(s)|^2, \tag{3.1.9}$$

where $f_{X,Y}^n(t, s) = \frac{1}{n}\Sigma_{k=1}^n \exp\{i\langle t, X_k\rangle + i\langle s, Y_k\rangle\}$, $f_X^n(t) = \frac{1}{n}\Sigma_{k=1}^n \exp\{i\langle t, X_k\rangle\}$ and $f_Y^n(s) = \frac{1}{n}\Sigma_{k=1}^n \exp\{i\langle s, Y_k\rangle\}$.

As the sample size increases, the distance dependence statistics converge almost surely to their respective population distance dependence measure. That is,

$$\lim_{n\to\infty} \mathcal{V}_n(\mathbf{X}, \mathbf{Y}) = \mathcal{V}(X, Y),$$
$$\lim_{n\to\infty} \mathcal{R}_n^2(\mathbf{X}, \mathbf{Y}) = \mathcal{R}^2(X, Y).$$

Unlike the RV coefficient which is discussed in Section 3.4 and the Wilk's lambda [101], the $\mathcal{R}_n$ or dCor coefficient is able to capture nonlinear associations, since it is based on a characterization of independence. Thus, more complex associations can be detected since it is sensitive to all types of departures from independence, including nonlinear or nonmonotone dependence structure. More specifically, the properties of $\mathcal{R}$ or dCor are enumerated below:

1. $0 \leq \mathcal{R} \leq 1$, whenever X and Y have finite first moments.

2. $\mathcal{R}(X, Y) = 0$ if and only if X and Y are independent.

3. It is consistent: It converges almost surely to its population counterpart $\mathcal{R}$ as $n \to \infty$.

4. If $\mathcal{R}_n(\mathbf{X}, \mathbf{Y}) = 1$, then there exist a vector a, a real number b and an orthogonal matric C such that $\mathbf{Y} = a + b\mathbf{X}\mathbf{C}$.

5. $\mathcal{R}$ is invariant to all shift and orthogonal transformations of $\mathbf{X}$ and $\mathbf{Y}$.

6. $\mathcal{R}$ is scale invariant.

7. If $p = q = 1$ with Gaussian distribution: $\mathcal{R} \leq |\rho|$

$$\mathcal{R}^2 = \frac{\rho \arcsin(\rho) + \sqrt{(1-\rho^2)} - \rho \arcsin(\frac{\rho}{2}) - \sqrt{4-\rho^2} + 1}{1 + \frac{\pi}{3} - \sqrt{3}} \tag{3.1.10}$$

Another good property of dCor coefficient is that its associated test based on $n\mathcal{V}_n^2(X, Y)$ is able to test whether there is independence between the random vectors X and Y. Under the null hypothesis, Székely, Rizzo and Bakirov [91] showed that the normalized test statistic, given by

$$T = \frac{n\mathcal{V}_n^2}{S_2}, \tag{3.1.11}$$

where $S_2 = \frac{1}{n^2}\sum_{k,l=1}^{n} |X_k - X_l|_p \frac{1}{n^2}\sum_{k,l=1}^{n} |Y_k - Y_l|_q$, converges in distribution to a quadratic form:

$$Q = \sum_{j=1}^{\infty} \lambda_j Z_j^2,$$

where Z_j are independent standard Gaussian random variables and λ_j are nonnegative constants that depend on the distribution of (X, Y).

If $E(|X|_p + |Y|_q) < \infty$, then the following statements are true.

1. If X and Y are independent, $n\mathcal{V}_n^2/S_2 \to Q$ in distribution as $n \to \infty$.

2. If X and Y are dependent, $n\mathcal{V}_n^2/S_2 \to \infty$ in probability as $n \to \infty$.

Hence, the null hypothesis is rejected for large values of $n\mathcal{V}_n^2/S_2$. In addition, a test rejecting independence of X and Y when $\sqrt{n\mathcal{V}_n^2/S_2} \geq \Phi^{-1}(1-\alpha/2)$ has an asymptotic significance level at most α. The asymptotic test criterion could be very conservative for many distributions. However, the best feature of this test is that it is consistent against all dependent alternatives with finite absolute first moment whereas some alternatives are ignored in the test based on Wilk's Lambda and RV coefficient.

According to Székely and Rizzo [92, 91], distance covariance (dCov) and distance correlation (dCor) are analogous to Pearson's product-moment covariance and correlation but they generalize and extend these classical bivariate measures of dependence.

To implement the dCor coefficient and the dCov test of independence, Rizzo and Székely [77] developed an R package *energy* with the functions `dcor` and `dcov.test`. To test the significance of the distance correlation coefficient, permutation tests are utilized.

3.1.2 Distance Correlation in High Dimension

Székely and Rizzo [94] proposed an unbiased modified version of distance covariance and distance correlation, which is favorable as dimensions p, q tend to infinity. The modification of the squared distance covariance resulted to a $t-$test of multivariate independence applicable in high dimension. The resulting t-test is unbiased for every sample size greater than three and all significance levels.

Székely and Rizzo showed that $\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})$ is a biased estimator of $\mathcal{V}^2(X, Y)$, and the bias in $\mathcal{R}_n^2(\mathbf{X}, \mathbf{Y})$ increases with dimension. That is, as $p, q \to \infty$, the sample distance correlation $\mathcal{R}_n^2(\mathbf{X}, \mathbf{Y}) \to 1$ even though X and Y are independent. Therefore, they constructed a modified distance covariance statistic such that under independence, a transformation of its corresponding distance correlation statistic converges to a student's *t*-distribution as $p, q \to \infty$, which is approximately normal when $p, q > n \geq 10$. This $t-$distributed statistic imparts easy interpretation of the sample correlation coefficient for high dimensional data.

Definition 3.5. *The unbiased distance covariance statistic is*

$$\mathcal{V}_n^*(\mathbf{X}, \mathbf{Y}) = \frac{\mathcal{U}_n^*(\mathbf{X}, \mathbf{Y})}{n(n-3)} = \frac{1}{n(n-3)} \left(\sum_{i,j=1}^{n} A_{i,j}^* B_{i,j}^* - \frac{n}{n-2} \sum_{i=1}^{n} A_{i,i}^* B_{i,i}^* \right) \tag{3.1.12}$$

where

$$\mathcal{U}_n^*(\mathbf{X},\mathbf{Y}) = \sum_{i\neq j} A_{i,j}^* B_{i,j}^* - \frac{2}{n-2}\sum_{i=1}^{n} A_{i,i}^* B_{i,i}^*$$

and

$$A_{i,j}^* = \begin{cases} \frac{n}{n-1}(A_{i,j} - \frac{a_{ij}}{n}) \ , & i \neq j; \\ \frac{n}{n-1}(\bar{a}_i - \bar{a}), & i = j, \end{cases}$$

and

$$B_{i,j}^* = \begin{cases} \frac{n}{n-1}(B_{i,j} - \frac{b_{ij}}{n}) \ , & i \neq j; \\ \frac{n}{n-1}(\bar{b}_i - \bar{b}), & i = j. \end{cases}$$

$\mathcal{V}_n^*(\mathbf{X},\mathbf{Y})$ *is an unbiased estimator of the squared population distance covariance,* $\mathcal{V}^2(X,Y)$.

It was proved by Székely and Rizzo [94] that the modified distance variance $\mathcal{U}_n^*(\mathbf{X},\mathbf{X}) \geq 0$ and $\mathcal{U}_n^*(\mathbf{Y},\mathbf{Y}) \geq 0$ so it follows that $\sqrt{\mathcal{V}_n^*(\mathbf{X},\mathbf{X})\mathcal{V}_n^*(\mathbf{Y},\mathbf{Y})}$ is always a real number for $n \geq 3$.

Definition 3.6. *The modified distance correlation statistic is*

$$\mathcal{R}_n^*(\mathbf{X},\mathbf{Y}) = \begin{cases} \frac{\mathcal{V}_n^*(\mathbf{X},\mathbf{Y})}{\sqrt{\mathcal{V}_n^*(\mathbf{X},\mathbf{X})\mathcal{V}_n^*(\mathbf{Y},\mathbf{Y})}} \ , & \mathcal{V}_n^*(\mathbf{X},\mathbf{X})\mathcal{V}_n^*(\mathbf{Y},\mathbf{Y}) > 0; \\ 0, & \textit{otherwise.} \end{cases} \tag{3.1.13}$$

Although the original $\mathcal{R}_n$ is between 0 and 1, the modified dCor statistic $\mathcal{R}_n^*$ can take on negative values. The modified dCor statistic $\mathcal{R}_n^*$ converges to the square of population distance correlation ($\mathcal{R}^2$) stochastically.

The test statistic for independence in high dimension, which is based on an unbiased estimator of the distance covariance, is given by

$$\mathcal{T}_n = \sqrt{v-1} \cdot \frac{\mathcal{R}_n^*}{\sqrt{1-(\mathcal{R}_n^*)^2}}. \tag{3.1.14}$$

This test is unbiased for every sample size greater than or equal to four ($n \geq 4$) and any significance

level. As $p, q \to \infty$, $\mathcal{T}_n$ converges in distribution to Student t with $v-1$ degrees of freedom, where $v = n(n-3)/2$.

Székely and Rizzo [94] also obtained a $Z-$test of independence in high dimension. Their results show that under independence of X and Y, if the coordinates of X and Y are independent and identically distributed with positive finite variance, then the limit distribution of $(\mathcal{R}_n^* + 1)/2$ is a symmetric beta distribution with shape parameter $(v-1)/2$. Then it follows that in high dimension, the sampling distribution of $\sqrt{v-1}\mathcal{R}_n^*$ for $n \geq 10$ is approximately standard normal.

Rizzo and Székely [77] developed an R package *energy* with the functions `bcdcor` to compute for the bias-corrected dCor and `dcor.ttest` to implement the unbiased test for independence.

3.1.3 Unbiased Distance Correlation based on $\mathcal{U}-$centering

Another unbiased estimator of squared distance covariance $\mathcal{V}^2(X, Y)$, which is based on alternate type of double centering called $\mathcal{U}-$centering, is established by Székely and Rizzo [96]. It is formed in such a way that the essential properties of distance covariance are maintained. It is also algebraically equivalent to the unbiased estimator $\sqrt{\mathcal{V}_n^*(\mathbf{X}, \mathbf{Y})}$ defined in Subsection 3.1.2.

The original definition of distance covariance in (3.1.5) utilized a type of centering with versions A_{ij} and B_{ij} that have the property that all rows and columns sum up to zero. Another type of centering, denoted by $\tilde{A}_{ij}$ and $\tilde{B}_{ij}$, are called unbiased or $\mathcal{U}-$centering. The $\mathcal{U}-$centering has the additional property that all expectations are zero; that is, $E[\tilde{A}_{ij}] = 0$ for all i, j.

Definition 3.7. *Let $A = (a_{ij})$ and $B = (b_{ij})$ be symmetric, real-valued $n \times n$ matrices with zero diagonal and $n > 2$. Then the $(i,j) - th$ element of each of the $\mathcal{U}-$centered matrices $\tilde{A}$ and $\tilde{B}$ is given by*

$$\tilde{A}_{i,j} = \begin{cases} a_{i,j} - \frac{1}{n-2}\sum_{l=1}^{n} a_{i,l} - \frac{1}{n-2}\sum_{k=1}^{n} a_{k,j} + \frac{1}{(n-1)(n-2)}\sum_{k,l=1}^{n} a_{k,l}\ , & i \neq j; \\ 0, & i = j, \end{cases}$$

and

$$\tilde{B}_{i,j} = \begin{cases} b_{i,j} - \frac{1}{n-2}\sum_{l=1}^{n} b_{i,l} - \frac{1}{n-2}\sum_{k=1}^{n} b_{k,j} + \frac{1}{(n-1)(n-2)}\sum_{k,l=1}^{n} b_{k,l}\ , & i \neq j; \\ 0, & i = j, \end{cases}$$

respectively.

An unbiased estimator of squared distance covariance, defined by Székely and Rizzo, is generated by this type of centering. This new method of centering is advantageous in defining the partial distance correlation. Partial distance correlation are thoroughly discussed in the paper by Székely and Rizzo [96].

Definition 3.8. *Let $(x_i, y_i), i = 1, \ldots, n$ be a sample of observations from the joint distribution (X, Y) of random vectors X and Y. Also, let $A = (a_{ij})$ be the Euclidean distance matrix of the sample $x_1, \ldots, x_n$ from the distribution of X, and $B = (b_{ij})$ be the Euclidean distance matrix of the sample $y_1, \ldots, y_n$ from the distribution of Y. If $E(|X| + |Y|) < \infty$ and $n > 3$, then an unbiased estimator of the squared distance covariance $\mathcal{V}^2(X, Y)$ is the inner product of two $\mathcal{U}-$centered matrices $\tilde{A}$ and $\tilde{B}$; that is,*

$$(\tilde{A} \cdot \tilde{B}) := \frac{1}{n(n-3)} \sum_{i \neq j} \tilde{A}_{i,j} \tilde{B}_{i,j}. \tag{3.1.15}$$

It is clear in the definition that $\tilde{A} = 0$ if all of the sample observations are identical. Furthermore, $\tilde{A} = 0$ if and only if the n sample observations have equal distances from each other or at least $n - 1$ of the n sample observations are exactly alike.

Here, Székely and Rizzo define a Hilbert space $\mathcal{H}_n$ generated by Euclidean distance matrices of arbitrary sets of n points in a Euclidean space $\mathbb{R}^p$ with $p \geq 1$. They consider $A = (a_{ij})$ as an arbitrary element in $\mathcal{S}_n$, the linear span of all $n \times n$ distance matrices of samples $x_1, \ldots, x_n$ and $B = (b_{ij})$ as an arbitrary element in $\mathcal{S}_n$, the linear span of all $n \times n$ distance matrices of samples $y_1, \ldots, y_n$. Székely and Rizzo showed that the linear span of all $n \times n$ matrices $\mathcal{H}_n = \tilde{A} : A \epsilon \mathcal{S}_n$ is a Hilbert space with inner product defined by (3.1.15).

The population distance covariance $\mathcal{V}(X,Y)$ in (3.1.1) has been defined in terms of the joint and marginal characteristic functions of the random vectors. An equivalent definition is given below adopting Lyons [64] generalized idea of distance correlation in separable Hilbert spaces.

$$\mathcal{V}^2(X,Y) := E\{\hat{A}_X \hat{B}_Y\}$$

where $\hat{A}_X$ is the abbreviation of $\hat{A}_X(x,x')$, which corresponds to the double centering function with respect to X, and $\hat{B}_Y$ is the abbreviation of $\hat{B}_Y(y,y')$, the corresponding double centering function with respect to Y. Here $\hat{A}_X(x,x')$ is a real valued function of two realizations of X and the subscript X references the underlying random variable. Similarly for $\hat{B}_Y(y,y')$.

The corresponding double centering functions are defined as

$$\begin{aligned}\hat{A}_X(x,x') = a(x,x') - \int_{\mathbb{R}^p} a(x,x')dF_X(x') - \int_{\mathbb{R}^p} a(x,x')dF_X(x) \\ + \int_{\mathbb{R}^p}\int_{\mathbb{R}^p} a(x,x')dF_X(x')dF_X(x),\end{aligned}$$

and

$$\begin{aligned}\hat{B}_Y(y,y') = b(y,y') - \int_{\mathbb{R}^q} b(y,y')dF_Y(y') - \int_{\mathbb{R}^q} b(y,y')dF_Y(y) \\ + \int_{\mathbb{R}^q}\int_{\mathbb{R}^q} b(y,y')dF_Y(y')dF_Y(y),\end{aligned}$$

provided the integrals exist.

The definition utilizes the bivariate distance functions $a(x,x') = |x-x'|_p$ and $b(y,y') = |y-y'|_q$, where x, x' are realizations of the random variables X and y, y' are realizations of the random variables Y. The random versions are also considered. The random distance functions are defined as $a(X,X') = |X-X'|_p$ and $b(Y,Y') = |Y-Y'|_q$ where $X \epsilon \mathbb{R}^p$ and $Y \epsilon \mathbb{R}^q$ are random variables with finite expectations.

3.2 Global Gaussian Correlation

Local Gaussian correlation (LGC) is another new measure of dependence developed by Tjøstheim and Hufthammer [99]. It is used to construct not only a global measure and test but also local dependence, which makes it unique among the measures considered. The global Gaussian correlation (τ or GGC) aggregates local version of correlation on subsets of $\mathbb{R}^2$ with bivariate Gaussian distribution into a global measure of dependence.

The idea of LGC is to approximate the bivariate distribution by a family of Gaussian bivariate distributions using local likelihood. At each point of the distribution there is a Gaussian distribution that gives a good approximation at that point. The correlation of the approximating Gaussian distribution is taken as the local correlation in that neighbourhood. This results in a nonlinear dependence measure, which is inherently local. It is formally defined as follows: Given a density function $f(X, Y)$, approximate f locally by a bivariate Gaussian distribution ϕ; that is, at the point $x = (x, y)$, or in a neighbourhood, a bivariate Gaussian density is fitted by

$$\begin{aligned}\phi(\theta(x), v) = \frac{1}{2\pi\sigma_1(x)\sigma_2(x)\sqrt{1-\rho^2(x)}} exp\{-\frac{1}{1-\rho^2(x)}\{\frac{[v_1-\mu_1(x)]^2}{\sigma_1^2(x)} \\ - 2\rho(x)\frac{[v_1-\mu_1(x)]}{\sigma_1(x)}\frac{[v_2-\mu_2(x)]}{\sigma_2(x)} + \frac{[v_2-\mu_2(x)]^2}{\sigma_2^2(x)}\}\}\end{aligned} \tag{3.2.1}$$

where $v = (v_1, v_2)^T$ is the running variable, $\theta(x) = [\mu_1(x), \mu_2(x), \sigma_1^2(x), \sigma_2^2(x), \rho(x)]^T$ and $\rho(x)$ is the local correlation. Then a local mean and a local variance will be computed based on estimation by local likelihood. Generally, the estimation of local likelihood involves estimating a density function $f(x)$ by a known parametric family say, $g(x, \theta)$. In this case, the emphasis is on estimating $\theta(x)$ not on $f(x)$.

Given the observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, the ordinary log likelihood for a Gaussian density ϕ is given by

$$L = \frac{1}{n}\sum_{i}^{n} \log \phi(X_i, Y_i),$$

where $\phi(X, Y)$ is given in (3.2.1). The following maximum likelihood estimate of ρ is used:

$$\hat{\rho} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{(\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2)}}.$$

Tjøstheim et al. [99] introduced kernel functions $K_{h_1}(X_t - x)$ and $K_{h_2}(Y_t - y)$ to describe a neighborhood of A around (x, y). The kernel functions are defined as $K_{h_1} = h_1^{-1}K = h_1^{-1}(X_t - x)$ and similarly, $K_{h_2} = h_2^{-1}K = h_2^{-1}(Y_t - y)$, where h_1 and h_2 are the bandwidths in x and y directions, respectively. Then, the appropriate local likelihood associated with the distribution $\phi(X, Y)$ is

$$L' = \frac{1}{n}\sum_{i}^{n} K_{h_1}(X_t - x)K_{h_2}(Y_t - y)\log\phi(X_i, Y_i).$$

However, an adjustment is needed that results in the following local log likelihood:

$$L = \frac{1}{n}\sum_{i}^{n} K_{h_1}(X_t - x)K_{h_2}(Y_t - y)\log\phi(X_i, Y_i) - \int K_{h_1}(v_1 - x)K_{h_2}(v_2 - y)\phi(v_1, v_2)dv_1dv_2.$$

The adjustment made use of a type of penalty function that was used by Hjort and Jones [44] for density estimation purposes. They argued that it can be interpreted as a locally weighted Kullback-Leibler criterion for measuring the distance between $f(\cdot)$ and $\phi(\cdot, \theta)$.

Then, the local likelihood estimates $\hat{\theta}_{h,n}(x, y)$ satisfy the equations of $\partial L/\partial\theta_j = 0$, which is seen as

$$\begin{aligned}\partial L/\partial\theta_j = &\frac{1}{n}\sum_{i}^{n} K_{h_1}(X_t - x)K_{h_2}(Y_t - y)\log w_j(X_i, Y_i, \theta) \\ &- \int K_{h_1}(v_1 - x)K_{h_2}(v_2 - y)w_j(v_1, v_2, \theta)\phi(v_1, v_2, \theta)dv_1dv_2. \qquad (3.2.2)\end{aligned}$$

The resulting 5-dimensional set of equations are solved numerically. It produces an estimate for local correlation $\hat{\rho}_h(x, y)$, estimates for local means $\hat{\mu}_{1,h}(x, y)$, $\hat{\mu}_{2,h}(x, y)$, and estimates for local variances, $\hat{\sigma}^2_{1,h}(x, y)$, $\hat{\sigma}^2_{2,h}(x, y)$, which can then be used to obtain local covariances.

If $n \to \infty$ for fixed h_1 and h_2 and using the law of large numbers on the first term of (3.2.2), $\partial L/\partial\theta_j$ converges towards $\int K_{h_1}(v_1 - x)K_{h_2}(v_2 - y)w_j(v_1, v_2, \theta)[f(v_1, v_2) - \phi(v_1, v_2, \theta)]dv_1dv_2$. For small bandwidths using smoothing conditions, and requiring $\partial L/\partial\theta_j = 0$ for all j, and

$$w_j(x, y, \theta(x, y))[f(x, y) - \phi(x, y, \theta(x, y))] + O(h^T h) = 0$$

and the local likelihood estimates satisfying $\partial L/\partial\theta_j = 0$, restrict $\phi(v_1, v_2, \theta(x, y))$ to be close to $f(x, y)$ when (u, v) is close to (x, y). This is the reason that the family $\phi(x, y)$ approximates f as the neighborhood defined by the bandwidth $h = [h_1, h_2]$ shrinks.

In obtaining the estimates of the standard errors, two methods can be considered. The first one is the bootstrap method which is valid in the case when X_i consists of independent and identically distributed random variables and this was used by Berentsen and Tjøstheim [6]. But since for each bootstrap realization the local likelihood has to be optimized numerically, it was observed to be very time-consuming. So another method which makes use of Monte Carlo method is suggested. It is described thoroughly in the paper of Tjøstheim and Hufthammer [99]. The choice of the bandwidth, according to Berentsen and Tjøstheim [6], depends largely on the purpose of the user. If the user wants to describe the local dependence structure in the data, it is useful to compute the local correlation $\rho_{n,h}(x, y)$ for several bandwidths to know the dependence structure on the different scales of locality. But, they suggested that a data-driven choice of bandwidth like the bandwidth choice for density kernel estimation is better. Tjøstheim et al. [99] discussed an approach to choose the bandwidth, which involves a compromise between optimizing the bias reduction for a density estimate and the choice of the degree of the variance for a local correlation estimate. However, Berentsen et al. [6] said that this bandwidth algorithm is not really satisfactory in a general situation. They proposed a general method based on the principle of likelihood cross-validation. Detailed discussion can be found in their papers [99, 6, 5].

The following are the properties of Local Gaussian correlation:

1. Range:

$$-1 \leq \widehat{\rho}_h(x, y) \leq 1$$
$$-1 \leq \rho_h(x, y) \leq 1.$$

2. If f is Gaussian, $\rho_h(x, y) = \rho$ is constant.

3. With linear transformation,

$$X_i' = a_1 + b_1 X_i \qquad Y_i' = a_2 + b_2 Y_i$$
$$h_i' = b_i h_i \qquad \widehat{\rho}_{X',Y',h'}(x', y') = \widehat{\rho}_{X,Y,h}(x, y).$$

4. X, Y independent implies $\rho_h(x, y) \equiv 0$ but not vice versa. $\rho_h(x, y) \equiv 0$ implies independence only for Gaussian variables.

Bivariate densities f with different types of symmetries were also observed. It is assumed that $\mu = E(X) = 0$. Symmetry properties of $\Sigma(x)$ and $\mu(x)$ can conceivably be used to obtain more precise estimates. Berentsen and Tjøstheim [6] discussed that they are used to increase the power of independence tests.

1. Orthogonal symmetry: f is invariant to orthogonal transformations but not $\rho_h(x, y)$.

2. Radial symmetry: $\rho_h(-x, -y) = \rho_h(x, y)$

3. Odd symmetry:

$$\rho_h(-x, y) = -\rho_h(x, y)$$
$$\rho_h(x, -y) = -\rho_h(x, y)$$

4. Exchange symmetry: $\rho_h(x, y) = \rho_h(y, x)$

5. Rotations: It implies that as the diagonals are approached along the density contours, which are ellipses:

$$\rho_h(x,x) > 0,$$
$$\rho_h(x,-x) < 0,$$
$$\rho_h(x,0) = \rho_h(0,y) = 0.$$

The rotation matrix is $A = \left(\begin{smallmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{smallmatrix}\right)$. For a given vector x, Ax is rotated counterclockwise through an angle α. Considering an arbitrary spherical density f, then f is a rotation symmetric in addition to being radial, reflection and exchange symmetric. In addition, the local correlation is radial and exchange but it is odd reflection symmetric. Rotating from a point $x = (x_1, 0)$ on the positive x_1 axis, $\rho(x_1, 0) = 0$ and $\Sigma(x)$ is diagonal resulting to

$$y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = A \begin{bmatrix} x_1 \\ 0 \end{bmatrix} = \begin{bmatrix} x_1 \cos\alpha \\ x_1 \sin\alpha \end{bmatrix}$$

and

$$\Sigma(y) = A\Sigma(x)A^T = \begin{bmatrix} \sigma_1(x)\cos^2(\alpha) + \sigma_2(x)\sin^2(\alpha) & (\sigma_1(x) - \sigma_2(x))\sin\alpha\cos\alpha \\ (\sigma_1(x) - \sigma_2(x))\sin\alpha\cos\alpha & \sigma_1(x)\sin^2(\alpha) + \sigma_2(x)\cos^2(\alpha) \end{bmatrix}.$$

It implies that

$$\begin{aligned} \rho^2 &= \rho^2(\alpha) \\ &= \frac{(\sigma_1(x) - \sigma_2(x))^2}{\sigma_1^2(x) + \sigma_2^2(x) + \sigma_1(x)\sigma_2(x)\left(\tan^2\alpha + \frac{1}{\tan^2\alpha}\right)}, \end{aligned} \tag{3.2.3}$$

which is maximum when $\tan^2\alpha = 1$; that is, $\alpha = \pm\frac{\pi}{4}$. It can be observed that $\rho(\alpha) > 0$ in quadrant I and quadrant III and $\rho(\alpha) < 0$ in quadrant II and quadrant IV. This implies that $\rho^2(\alpha) > 0$ in all quadrants.

The advantages of LGC are summed up in the following statements. It gives a complete characterization of the dependence locally since it describes the dependence relation for a function f at each point. It does not suffer from the bias problem caused by the conditional correlation. Another is that it is able to detect and quantify more complex, nonlinear changes in the dependence structure as well as capable of detecting asymmetric dependence.

Berentsen and Tjøstheim [6] constructed a global measure of dependence by aggregating the local correlations $\rho(x, y)$ on subsets of $\mathbb{R}^2$. They considered the functional $\rho^2(x, y)$ to avoid the cancellation of local correlation in different points because for a nonlinear dependence structure, the local Gaussian correlation can have positive or negative signs.

Definition 3.9. *The global measure of dependence is defined as*

$$\tau = (E_F(\rho^2(X, Y)))^{1/2} = \left(\int \rho^2(x, y) dF(x, y) \right)^{1/2} \tag{3.2.4}$$

where $F(x, y)$ is the joint distribution function of X and Y.

The properties of the global Gaussian correlation (τ or GGC) are the following:

1. Range: $0 \leq \tau \leq 1$
2. Independence: If X and Y are independent, then $\tau = 0$.
3. Functional dependence: For any Borel-measurable function g, if $Y = g(X)$ (or vice versa), then $\tau = 1$.
4. Gaussian case: If the joint distribution of X and Y is Gaussian with correlation coefficient ρ then $\tau \equiv |\rho|$.

Definition 3.10. *The statistic of the global measure of dependence is defined by*

$$\tau_{n,h} = (E_{F_n}(\rho^2_{n,h}(X, Y)))^{1/2} = \left(\int \rho^2_{n,h}(x, y) dF_n(x, y) \right)^{1/2}, \tag{3.2.5}$$

where $F_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} 1(X_i \leq x, Y_i \leq y)$ with $1(\cdot)$ denoting the indicator function.

Another version of the sample global measure of dependence that protects outliers outside some subset S of $\mathbb{R}^2$ is given in the next equation. The scaling $\int 1_S(x,y)dF_n(x,y)$ is done to make sure that $0 \leq \tau_{n,h}(S) \leq 1$.

$$\tau_{n,h}(S) = \frac{(\int \rho^2_{n,h}(x,y)1_S(x,y)dF_n(x,y))^{1/2}}{\int 1_S(x,y)dF_n(x,y)}. \tag{3.2.6}$$

Moreover, the asymptotic properties of the statistics as shown in [99] give the following results:

1. $\hat{\theta}_h(x,y)$ converges in distribution to $\theta_h(x,y)$ when h is fixed and n tends to infinity.

2. $\hat{\theta}_{h,n}(x,y)$ is asymptotically normal such that

$$(nh_1h_2)^{1/2}J_hM_h^{1/2}[\hat{\theta}_{h,n}(x,y) - \theta_h(x,y)] \to \mathcal{N}(0,I)$$

   where I is the identity matrix of dimension 5 and J_h and M_h are defined in [6, 99, 44].

3. $T_{n,h} \to T$ which means that $T_{n,h}$ is a consistent test statistic, where

$$T_{n,h} \doteq \int_S g(\rho_{n,h}(x,y))dF_n(x,y)$$

   estimates the functional $T \doteq \int_S g(\rho(x,y))dF(x,y)$ and $g(x)$ are any variance stabilizing functions but here, $g(x) = x$ and $g(x) = x^2$ were used.

4. $\sqrt{n}(T_{n,h} - T_h) \to \mathcal{N}(0, \int A_h(x)^2dF(x) - \int\int A_h(x)A_h(y)\,dF(x)\,dF(y)$ where $A_h(x)$ and $A_h(y)$ are defined in [6].

An R-package *localgauss* has been developed by Berentsen et al. [5] for finding the local likelihood estimates $\theta_{n,h}(x)$ including $\rho_{n,h}(x)$. It is available publicly at the Comprehensive R Archive Network for Linux and Windows [69]. In testing the hypothesis

$$H_0\text{: } X \text{ and } Y \text{ are independent}$$

vs

H_1: X and Y are not independent,

bootstrap method and permutation test were used because the asymptotic theory for functionals of type $T_{n,h}$ is not accurate unless n is very large. The procedures are detailed by Berentsen and Tjøstheim [6]. The implementation is included in the Appendix.

3.3 Maximal Information Coefficient

Maximal Information Coefficient (MIC), introduced by Reshef, Reshef, Finucane, Grossman, McVean, Turnbaugh, Lander, Mitzenmacher and Sabeti [75], is an exploratory data analysis tool that begins by exploring a large data set through searching pairs of variables that are closely associated. A measure of dependence is calculated for each pair, pairs are ranked by scores and then top-scores examined. In short, MIC summarizes the grid of pairwise dependencies in a large set of variables. Reshef et al. stated that this works when two heuristic properties are satisfied: generality and equitability. These properties are defined below. More discussion on the property of equitability is found in Chapter 4.

Reshef et al. [75] illustrate that MIC belongs to a larger class of maximal information-based nonparametric exploration (MINE) statistics, which is not only used for identifying interesting relationships but classifying them according to properties such as nonlinear and monotonicity.

According to Reshef et al. [75], MIC is based on the idea that if there exists a relationship between two variables, then this relationship can be reflected on a grid drawn on the scatterplot that partitions the data to encapsulate that relationship. Thus, the MIC of a bivariate data is computed using the following algorithm:

1. Explore all grids up to a maximal grid resolution depending on the sample size n.

2. For every pair of integers (x, y), compute the largest possible mutual information (MI) achievable by any x-by-y grid applicable to the data. A technique called binning method is used such that if close values of $\mathbf{X}$ are grouped together and close values of $\mathbf{Y}$ are also grouped together then MI can be used. The bins are considered as the random variables.

They are chosen in such a way that the MI is maximal when $H(X_b) = H(Y_b) = H(X_b, Y_b)$. The $I(D|_G)$ denote the mutual information of the probability distribution induced on the boxes of grid G, where the probability of a bin is proportional to the number of data points falling inside the bin.

3. Normalize the mutual information values to ensure a fair comparison between grids of different dimensions and to obtain modified values between 0 and 1.

4. Obtain the highest normalized mutual information achieved by any x-by-y grid. Therefore, MIC is the maximum value of all the highest normalized mutual information over ordered pairs (x, y) such that $xy < B$ where B is a function of sample size which is usually set as $B = n^{0.6}$ or $B = n^{0.55}$.

Reshef et al. [75] gives the following definition of MIC:

Definition 3.11. *Let D be a finite set of ordered pairs. For a grid G, let $D|_G$ denote the probability distribution induced by the data D on the cells of G, and let $I(\cdot)$ denote mutual information. Let $I^*(D, x, y) = max_G I(D|_G)$, where the maximum is taken over all x-by-y grids G (possibly with empty rows/columns). Then*

$$MIC(D) = max_{xy<B(|D|)} \frac{I^*(D, x, y)}{\log_2 \min\{x, y\}}, \tag{3.3.1}$$

where B is the growing function satisfying $B(n) = O(n)$, and

$$I(X, Y) = \sum_x \sum_y p(x, y) \log\{\frac{p(x, y)}{p(x)p(y)}\}, \tag{3.3.2}$$

with $p(x, y)$ as the joint probability distribution of x and y, $p(x)$ and $p(y)$ are the marginal distributions of x and y, respectively.

In particular, Reshef et al. [75] referred to the ratio in the definition as a characteristic matrix

$M(D)$, that is,

$$M(D) = \frac{I^*(D, x, y)}{\log_2 \min\{x, y\}}. \tag{3.3.3}$$

Thus, MIC can also be written as

$$MIC(D) = max_{xy<B(|D|)} M(D). \tag{3.3.4}$$

According to Reshef et al. [75], the space of grids that must be explored to compute each entry $m_{x,y}$ of the characteristic matrix grows exponentially with the number of data points, so they use a heuristic dynamic programming algorithm to approximate MIC in practice for efficiency purposes.

The following are the properties of MIC statistic:

1. MIC falls between 0 and 1.
2. MIC is symmetric, that is, $MIC(X, Y) = MIC(Y, X)$.
3. MIC is invariant under order-preserving transformations of the x and y values of D since I_G depends only on the rank order of the data.
4. With probability approaching 1 as sample size increases,
   - MIC assigns scores that tend to 1 to all never-constant noiseless functional relationships.
   - MIC assigns scores that tend to 1 for a larger class of noiseless including superpositions of noiseless functional relationships.
   - MIC assigns scores that tend to 0 to statistically independent variables.
5. MIC satisfies the properties of generality and equitability. By generality, Reshef et al. [75] mean that with sufficient sample size, the statistic is able to capture a wide range of interesting associations not only limited to specific function types such as linear, exponential or periodic but also to all functional relationships. By equitability, they mean that the

statistic should give similar scores to equally noisy relationships of different types. Reshef, Reshef, Mitzenmacher and Sabeti [76] claim that equitability is important for analyzing high-dimensional data sets .

As observed, MIC is based on mutual information but according to Reshef et al. [76], it is not itself a mutual information estimator and direct estimation of mutual information does not yield an equitable statistic. Evidence of the claim was provided by comparing the estimates of MIC and mutual information using a range of data set sizes and noise models with different parameters. When using the direct mutual information estimation, they used the squared Linfoot correlation through the well-known estimator of Kraskov, Stogbauer and Grassberger [55] with two smoothing parameters ($k = 1$ and $k = 6$), to normalize the resulting scores to obtain a measure between 0 and 1. It was found that mutual information is significantly less equitable than MIC across all the noise models tested when the sample size is $n = 500$. Likewise, at $n = 5000$, MIC outperforms mutual information on the most noise models except for those with vertical noise setting, where they behave similarly.

According to Reshef et al. [76], the results of the simulation analysis illustrate that the maximization and the normalization in the definition of MIC are necessary for its equitability. Without the normalization step, relationships that are better captured by grids with more cells are favored over those that are better captured by simple grids. Without the maximization step, relationships that are not naturally equipartitioned are unduly penalized.

Moreover, Reshef et al. [75] admitted that there is a tradeoff between equitability and power in the MIC statistic. They compared MIC with distance correlation, an elegant measure of dependence based on Euclidean distances developed by Székely, Rizzo and Bakirov [91] discussed in Section 2.2. Distance correlation has a better power than MIC for many relationship types but just like the classical Pearson product-moment correlation, it is highly non-equitable across all noise models tested. MIC, as a tool for data exploration, is able to capture the strongest relationships in a data set, but it was found out by Simon and Tibshirani [85] and Gorfine, Heller and Heller [36] that MIC has lower power than other methods such as distance correlation for detecting as many

weak relationships as possible.

There are some disadvantages of MIC that Kinney and Atwal [54] discovered. First, MIC is completely insensitive to certain types of noise. Second, MIC is not invariant under nonmonotonic transformations of X and Y. They also listed several arguments that disprove the findings of Reshef et al.

1. The definition of MIC could result to over fitting, and the choice of the parameter $B = n^{0.6}$ or $B = n^{0.55}$ in the definition does not have a mathematical proof.

2. The mathematical definition of equitability provided by Reshef et al. cannot be satisfied by any (nontrivial) dependence measure.

3. The simulation evidence presented by Reshef et al. are inaccurate.

4. Mutual information being less equitable than MIC is faulty.

An anonymous writer [33] at Gelman's blog states that the real problem with MIC is that an optimal grid is needed. He added that in many cases such an optimal grid simply does not exist at all when the MIC value increases with the number of bins, and hence the reported MIC value corresponds to the (user specified) maximum grid size. Increasing this maximum grid size will affect the values for different pairs of variables differently then giving any comparisons a slightly negative connotation.

The computation of MIC is possible using the Maximal Information-based Nonparametric Exploration (MINE) suite of Reshef et al. [75] which can be found at the MINE website [74]. The MINE authors provided a Java implementation (MINE.jar) and two wrappers (R and Python) for computation. There is also a *minerva* package written by Filosi, Visintainer, Albanese, Riccadonna, Jurman, and Furlanello [24] which provides the `mine` function that allows the computation of MINE statistics, including MIC. This package is an R wrapper for the C engine *cmine*. There are pre-computed p-values of various MIC scores at different sample sizes that can be found under *Downloads* in the MINE website [74]. According to Reshef et al. [75], the uncorrected

p-value of a given MIC score under a null hypothesis of statistical independence depends only on the score and on the sample size since MIC is a rank-order statistic.

3.4 RV Coefficient

The RV coefficient, which was developed by Escoufier [21], is a multivariate generalization of the squared Pearson correlation coefficient. It measures how two sets of points represented in matrices, X and Y, are far apart from each other. Like the simple Pearson correlation coefficient, it makes use of the concepts of covariance and variance, but of vector-valued random variables. Escoufier [22] defined RV coefficient as a similarity coefficient between positive semi-definite matrices.

Definition 3.12. *The population RV correlation coefficient or ρV of two random vectors X and Y is defined as*

$$\begin{aligned} \rho V(X,Y) &= \frac{CovV(X,Y)}{\sqrt{VaV(X)VaV(Y)}} \\ &= \frac{tr(\Sigma_{XY}\Sigma_{YX})}{\sqrt{tr(\Sigma_{XX}^2)tr(\Sigma_{YY}^2)}} \\ &= \frac{< XX', YY' >}{|XX'||YY'|} \\ &= \frac{tr(XX'YY')}{\sqrt{tr(XX')^2 tr(YY')^2}} \end{aligned} \tag{3.4.1}$$

where $CovV$ denotes the scalar-valued covariance and VaV denotes the scalar-valued variance.

According to Josse and Holmes in [51], the notion of RV coefficient is to consider that two sets of variables are correlated if the relative position of the samples in one set is similar to the relative position of the samples in another set.

Consider two random vectors X in $\mathbb{R}^p$ and Y in $\mathbb{R}^q$ and data matrices $\mathbf{X}_{n,p}$ and $\mathbf{Y}_{n,q}$ that are n independent realizations of the random vectors. The relative positions of the samples are represented by the cross-product of the matrices defined as: $\mathbf{XX}'$ and $\mathbf{YY}'$. The measure of

closeness between these samples is computed as the inner product:

$$< \mathbf{XX}', \mathbf{YY}' > = tr(\mathbf{XX}'\mathbf{YY}') \\ = \sum_{l=1}^{p}\sum_{m=1}^{q} cov^2(X_{.l}, Y_{.m}). \tag{3.4.2}$$

The RV statistic can also be written in terms of distance matrices:

$$RV(\mathbf{X}, \mathbf{Y}) = \frac{< C\Delta_{\mathbf{X}}^2 C, C\Delta_{\mathbf{Y}}^2 C >}{|C\Delta_{\mathbf{X}}^2 C||C\Delta_{\mathbf{Y}}^2 C|}, \tag{3.4.3}$$

where $C = I_n - \frac{1_n 1_{n'}}{n}$ with I_n the identity matrix of order n and 1_n a vector of ones of size n and Δ_{nxn} is the matrix where element d_{ij} represents the Euclidean distance between the samples i and j. This definition of RV uses the relationship of the inner product and the Euclidean distance between two samples as stated by Gower [37].

As defined by Robert and Escoufier [79], RV statistic is also equivalent to the following:

$$RV(\mathbf{X}, \mathbf{Y}) = \frac{tr(\mathbf{XX}'\mathbf{YY}')}{\sqrt{tr(\mathbf{XX}')^2 tr(\mathbf{YY}')^2}} \tag{3.4.4}$$

$$= \frac{tr(S_{\mathbf{XY}} S_{\mathbf{YX}})}{\sqrt{tr(S_{\mathbf{XX}}^2) tr(S_{\mathbf{YY}}^2)}} \tag{3.4.5}$$

where $S_{\mathbf{XY}} = \mathbf{X}'\mathbf{Y}$ which is the empirical covariance matrix between $\mathbf{X}$ and $\mathbf{Y}$.

The main properties of the RV correlation statistic are:

1. $0 \leq RV(X, Y) \leq 1$
2. $RV(X, Y) = 0$ if and only if $X'Y = 0$, that is, all the variables of one group are orthogonal to all the variables in the other group.
3. $RV(X, aBX + c) = 1$ where a is a constant, B an orthogonal matrix and c a constant vector.

4. When $p = q = 1$, $RV = r^2$ where r^2 is the square of the simple correlation coefficient.

5. RV is consistent, that is, when $n \to \infty$, it converges to its population counterpart ρV.

Like the simple Pearson product-moment correlation ρ for the univariate test, the statement holds for the multivariate case that $\rho V = 0$ does not necessarily imply that X and Y are independent unless an assumption of multivariate normality is satisfied. When $\rho V = 0$, it means that there is no linear relationship between X and Y. The hypothesis test of RV coefficient to assess the significance of the association is:

$$H_0 : \rho V = 0$$

$$H_1 : \rho V > 0$$

The hypothesis H_0 states that there is no linear association between the two sets X and Y, while the alternative hypothesis H_1 states that there is a linear association between the two sets X and Y.

The asymptotic distribution of the statistic nRV is available when the random variables have a multivariate normal joint distribution (Robert, Cléroux and Ranger [78]) or when it belongs to the class of elliptical distributions (Cléroux and Ducharme [15]). In those cases, nRV converges to:

$$\frac{1+k}{tr(\sum_{XX}^2)tr(\sum_{YY}^2)} \sum_{l=1}^{p} \sum_{m=1}^{q} \lambda_l \gamma_m Z_{lm}^2, \tag{3.4.6}$$

where k is the kurtosis parameter of the elliptical distribution, $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ are the eigenvalues of the covariance matrix Σ_{XX}, $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_q$ are the eigenvalues of the covariance matrix Σ_{YY}, and Z_{lm} are independent and identically distributed $N(0,1)$ random variables. The null hypothesis $H_0 : \rho V = 0$ (equivalent to $H_0 : \Sigma_{XY} = 0$) is rejected at level α, if $nRV > c_\alpha$ where c_α is the $(1-\alpha)$ percentile of the asymptotic distribution given in Equation (3.4.6).

If the distributional assumptions are not met, Cléroux, Lazraq and Lepage [16] suggested a test based on ranks, but Josse, Pagès and Husson [50] showed that this test is accurate only for sample

size $n > 300$.

A permutation test may be possible with RV coefficient, but there is a need to carefully construct the implementation since this is not equivalent to a complete permutation test of the vectorized cross-product matrices for which the exhaustive distribution is much larger.

It was found by Josse and Holmes [51] that computing the exact permutation distribution is computationally costly when $n > 15$. Hence, it is usually approximated by Monte Carlo Method, although a moment matching approach can also be used. The moment approach does not perform permutations but utilizes an analytical moment of the exact permutation distribution under the null. Kazi-Aoual, Hitier, Sabatier and Lebreton [52] defined the first moments of the quantity (3.4.2) under the null which outputs the moments of the RV coefficient. The expectation provides information of the expected behavior of the RV coefficient and is given by:

$$\mathbb{E}_{H_0}(RV) = \frac{\sqrt{\beta_x \times \beta_y}}{n-1},$$

where $\beta_x = (tr(X'X))^2/tr((X'X)^2)$ and $\beta_y = (tr(Y'Y))^2/tr((Y'Y)^2)$.

It can be seen that under the null, the RV coefficient takes on high values when the sample size is small and when $\mathbf{X}$ and $\mathbf{Y}$ are highly multi-dimensional. This is true since β_x is a measure of the complexity of the matrix. It varies between 1 and p: 1 when all the variables are perfectly correlated and p when all the variables are orthogonal.

Escoufier [79] further explained that RV coefficient is a unifying tool for linear multivariate statistical methods including principal component analysis, discriminant analysis and correlation analysis, to name a few. He considered that a common geometrical representation of a sample of p numerical variables with n observations arranged into a $p \times n$ matrix $\mathbf{X} = (x^{ij})$ consists of a canonical mapping of the data matrix $\mathbf{X}$ into a configuration of n points in the p-dimensional space $\mathbb{R}^p$. The behavior of such a configuration or the set of distances between its points is the basis of this measure. By using a positive semi-definite matrix Q, the distance between the j_{th} and the k_{th} points is equal to $\{(X^j - X^k)'Q(X^j - X^k)\}^{1/2}$. For any Q, there is a matrix L with

dimension $p \times q$ such that $Q = LL'$. This means that there is an equivalence between the choice of the metric defined by Q on points in $\mathbb{R}^p$ and the linear change of variables producing a new data matrix $\mathbf{Y} = L'\mathbf{X}$ followed by the use of the ordinary sum of squares metric on the configuration representing Y in $\mathbb{R}^q$.

There are R packages that implement the RV coefficient: *FactoMineR* [47] and *ade4* [20]. The *FactoMineR* used in this paper has the function `coeffRV` that provides computation of the measure and the Pearson type III approximation to test its significance.

3.5 Heller-Heller-Gorfine Statistics

Heller, Heller and Gorfine [42] developed a powerful nonparametric multivariate test of association which is applicable in all dimensions and is consistent against all alternatives. The test relies on norm-based distance metrics in X and (separately) in Y.

The problem is to test whether a relationship occurs between X and Y. The hypotheses are as follows:

H_0: $F_{XY} = F_X F_Y$ which means that the two vectors X and Y are independent.

vs

H_1: $F_{XY} \neq F_X F_Y$ which means that the two vectors X and Y are dependent.

The test statistic is a function of the ranks of the pairwise distances between the sample values of X and the sample values of Y. These distances are given by:

$$\{d_X(x_i, x_j) : i, j \epsilon 1, \ldots, N\}\{d_Y(y_i, y_j) : i, j \epsilon 1, \ldots, N\},$$

where $X \epsilon \mathbb{R}^p$ and $Y \epsilon \mathbb{R}^q$. In this setting, it is supposed that there are N independent copies of (x_i, y_i) for $i = 1, \ldots, N$ from the joint distribution of X and Y for testing H_0. The number of variables p, q can be greater than N since it works for all dimensions. The distance metrics $d_X(\cdot, \cdot)$ and $d_Y(\cdot, \cdot)$ that are considered here are restricted to those that are determined by norms. The test statistic is expressed in a simple closed form and is easy to implement.

As explained by Heller et al. [42], the test is motivated by the idea that if X and Y have a continuous joint density which are dependent, then there exists a point (x_0, y_0) in the sample space of (X, Y) and radii R_{x_0} and R_{y_0} around x_0 and y_0, respectively, such that the joint distribution of X and Y is different from the product of the marginal distributions in the Cartesian product of balls around (x_0, y_0).

Thus, Heller et al. defined $d(x_i, x_j)$ as the distance between two vectors x_i and x_j from the distribution of X and $d(y_i, y_j)$ the distance between two vectors y_i and y_j from the distribution of Y. So that in general, $d(\cdot\,, \cdot\,)$ is the norm distance between two sample points in either X or Y. Assuming that a point (x_0, y_0) and the radii R_{x_0} and R_{y_0} are known, they considered the dichotomous random variables

$$I\{d(x_0, X) \leq R_{x_0}\},$$
$$I\{d(y_0, Y) \leq R_{y_0}\},$$

where $I(\cdot\,)$ is the indicator function. Then, the observed cross-classification for the N independent observations are summarized in Table 3.1 below, where A_{ij}'s are defined as follows:

$$A_{11} = \sum_{k=1}^{N} I\{d(x_0, x_k) \leq R_{x_0}\} I\{d(y_0, y_k) \leq R_{y_0}\},$$
$$A_{12} = \sum_{k=1}^{N} I\{d(x_0, x_k) \leq R_{x_0}\} I\{d(y_0, y_k) > R_{y_0}\},$$
$$A_{21} = \sum_{k=1}^{N} I\{d(x_0, x_k) > R_{x_0}\} I\{d(y_0, y_k) \leq R_{y_0}\},$$
$$A_{22} = \sum_{k=1}^{N} I\{d(x_0, x_k) > R_{x_0}\} I\{d(y_0, y_k) > R_{y_0}\}.$$

and $A_{m\cdot}$ and $A_{\cdot m}$ for $m = 1, 2$ denote the row and column sums, respectively.

When the values of (x_0, y_0) as well as R_{x_0} and R_{y_0} are unknown, the data can be used to determine them. Each sample point will be considered in its turn to be (x_0, y_0) and every sample

Table 3.1: The cross-classification of $I\{d(x_0, X) \leq R_{x_0}\}$ and $I\{d(y_0, Y) \leq R_{y_0}\}$

Case	$d(y_0, \cdot) \leq R_{y_0}$	$d(y_0, \cdot) > R_{y_0}$	Row Total
$d(x_0, \cdot) \leq R_{x_0}$	A_{11}	A_{12}	$A_{1\cdot}$
$d(x_0, \cdot) > R_{x_0}$	A_{21}	A_{22}	$A_{2\cdot}$
Column Total	$A_{\cdot 1}$	$A_{\cdot 2}$	N

Table 3.2: The cross-classification of $I\{d(x_i, X) \leq d(x_i, x_j)\}$ and $I\{d(y_i, Y) \leq d(y_i, y_j)\}$

Case	$d(y_i, \cdot) \leq d(y_i, y_j)$	$d(y_i, \cdot) > d(y_i, y_j)$	Row Total
$d(x_i, \cdot) \leq d(x_i, x_j)$	$A_{11}(i,j)$	$A_{12}(i,j)$	$A_{1\cdot}(i,j)$
$d(x_i, \cdot) > d(x_i, x_j)$	$A_{21}(i,j)$	$A_{22}(i,j)$	$A_{2\cdot}(i,j)$
Column Total	$A_{\cdot 1}(i,j)$	$A_{\cdot 2}(i,j)$	$N-2$

point $j \neq i$ is used to define $R_{x_0} = d(x_i, x_j)$ and $R_{y_0} = d(y_i, y_j)$. Then, the 2×2 contingency tables, which are summarized in Table 3.2, contain the remaining $N-2$ points with $I\{d(x_i, X) \leq d(x_i, x_j)\}$ and $I\{d(y_i, Y) \leq d(y_i, y_j)\}$ considered for fixed observations i and j. The test collects all the evidence against independence by getting the sum over all $N(N-1)$ test statistics produced from the 2×2 contingency tables.

Table 3.2 shows the observed cross-classification for the $N-2$ independent observations $k \epsilon \{1, \ldots, N\}$ with $k \neq i, j$, where $A_{11}(i,j) = \sum_{k=1, k\neq i, k\neq j}^{N} I\{d(x_i, x_k) \leq d(x_i, x_j)\} I\{d(y_i, y_k) \leq d(y_i, y_j)\}$, A_{12}, A_{21}, A_{22} are similarly defined and $A_{m\cdot}$ and $A_{\cdot m} (m = 1, 2)$ denote the row and column totals.

Definition 3.13. *The test for independence of X and Y that Heller et al. [42] developed is given by:*

$$T = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} S(i,j), \tag{3.5.1}$$

where

$$S(i,j) = \frac{(N-2)\{A_{12}(i,j)A_{21}(i,j) - A_{11}(i,j)A_{22}(i,j)\}^2}{A_{1\cdot}(i,j)A_{2\cdot}(i,j)A_{\cdot 1}(i,j)A_{\cdot 2}(i,j)}$$

is the classic test statistic associated with Pearson's test for 2×2 contingency tables. $S(i,j)$ is set to 0, for i and j, if 0 is in at least one of the margins. The p-value from the permutation test based on the statistic T is the fraction of replicates of T under random permutations of the indices of the Y sample that are at least as large as the observed statistic.

The test is applicable for any dimensions and is consistent against all alternatives. It is consistent for both discrete and continuous random variables. It is a powerful test which has a simple form, easy to implement and has a good power.

Regarding computations, a naive implementation of the test will require $O(N^3)$ operations for N sample points since the remaining $N-2$ points are used for computing $S(i,j)$ for each pair (i,j). T is calculated efficiently in $O(N^2 \log N)$ time by using an algorithm, which for a given i, calculates $\{S(i,j) : j = 1,\ldots,N, j \neq i\}$ in $O(N \log N)$ operations.

The following steps are performed for HHG test:

1. Renumber the indices of the $N-1$ sample points according to increasing distance in X from i. In this way, the j^{th} observation is the j^{th} nearest to i in X.

2. Compute $\{\pi(j) : j = 1,\ldots,N, j \neq i\}$ where $\pi(1)\cdots\pi(N-1)$ are the ordered distances from i in Y, so that the j^{th} observation is the $\pi(j)^{th}$ nearest to i in Y. Here, $\pi(\cdot)$ is a permutation of $1,\ldots,N-1$

3. Compute $\{inv(j) : j = 1,\ldots,N, j \neq i\}$ where $inv(j)$ is the number of inversions of j in the permutation π, that is, the number of indices $k\epsilon\{1,\ldots,j-1\}$ such that $\pi(k)\epsilon\{\pi(j)+1, ldots, N-1\}$. From the definition of $A_{12}(i,j)$, the following are true:

$$A_{12}(i,j) = inv(j) \qquad A_{22}(i,j) = N-1-\pi(j)-inv(j)$$
$$A_{11}(i,j) = j-1-inv(j) \qquad A_{21}(i,j) = \pi(j)+inv(j)-j$$

since $A_{1\cdot}(i,j) = j-1$.

HHG test statistics and the p-values, using specified random permutations, are computed using the function `hhg.test` implemented in the *HHG* package [14] in R.

3.6 Pearson Product Moment Correlation

The Pearson product-moment correlation is the most widely known and applied correlation coefficient, developed by Karl Pearson [67]. It is used to measure the linear association between

two continuous random variables X and Y. The definition and properties included here were taken from Liebetran [62].

Definition 3.14. *The population coefficient of Pearson product-moment correlation is defined as*

$$\begin{aligned} \rho &= \frac{cov(X,Y)}{\sigma_X \sigma_Y} \qquad (3.6.1) \\ &= \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \\ &= \frac{E[XY]-E[X]E[Y]}{\sqrt{(E[X^2]-E[X]^2)(E[Y^2]-E[Y]^2)}} \end{aligned}$$

(3.6.2)

where $cov(X,Y)$ is the covariance of X and Y, σ_X is the standard deviation of X, μ_X is the mean of X and E is the expectation.

Definition 3.15. *The sample coefficient of Pearson product-moment correlation, or simply the sample correlation coefficient, is defined as*

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2 \sum_{i=1}^{n}(y_i-\bar{y})^2}}, \qquad (3.6.3)$$

where $x_i, i = 1, \ldots, n$ are values in the first data set, $y_i, i = 1, \ldots, n$ are values in the second data set, $\bar{x}$ is the mean of the values of x_i.

The properties of Pearson product-moment correlation are the following:

1. $-1 \leq \rho \leq 1$.

2. $\rho(X,Y) = \rho(Y,X)$.

3. If X and Y are independent, then $\rho = 0$. The statement $\rho = 0$ implies independence only when the joint distribution of X and Y is normal.

4. $\rho(X,Y)$ is invariant to separate changes in location and scale transformations.

(a) $\rho(X + a, Y + b) = \rho(X, Y)$

(b) $\rho(aX, bY) = \rho(X, Y)$

Details are shown in Section 4.2.2.

If the underlying variables X and Y have a bivariate normal distribution, then the statistic that is used to test the null hypothesis of independence is

$$t = r\sqrt{\frac{n-2}{1-r^2}}. \tag{3.6.4}$$

This statistic t has a Student's t-distribution with $v = n - 2$ as described by Rahman [70].

We include an investigation of Pearson correlation's properties, particularly the equitability and the rigid motion invariance in Section 4.2 and compare this classical coefficient with the more recent ones.

3.7 Maximal Correlation

Maximal correlation is a measure of association between two random variables X and Y that is developed by Gebelein [30]. It is frequently referred to by statisticians when correlation is discussed. It is known as the *sup* correlation defined by

$$\rho'(X, Y) = sup\left(\rho[f(X), g(Y)]\right), \tag{3.7.1}$$

where ρ represents the Pearson product-moment correlation coefficient, and supremum (sup) is taken over all Borel-measurable functions f and g for which $Var[f(X)]$ and $Var[g(X)]$ are finite and nonzero.

The properties of maximal correlation are enumerated by Balakrishnan and Lai [4]:

1. $0 \leq \rho'(X, Y) \leq 1$.

2. $\rho'(X, Y) = \rho'(Y, X)$.

3. $\rho'(X, Y) = 0$ if and only if X and Y are independent.

4. If X and Y are mutually dependent, then $\rho'(X, Y) = 1$ but the converse is not true as shown by Lancaster. [57].

5. $|\rho(X, Y)| \leq \rho'(X, Y)$

6. If (X, Y) is a bivariate normal random variable, $\rho'(X, Y) = |\rho(X, Y)| = |\rho|$.

It was proved by Rényi [73] that even if X and Y are only implicitly dependent, $\rho'(X, Y)$ is still equal to 1.

Balakrishnan et al. [4] stated that maximal correlation has many good properties, however, Hall [41] said that it has a lot of drawbacks. For example, the maximal correlation is often equal to 1 and is also computationally difficult. Hence, the maximal correlation coefficient has been less applied. But recently, there are new procedures established for computing maximal correlation and they are presented below.

Yenigün [103] discussed a method that followed from Rényi's [72] approach for estimating the maximal correlation for 2×2 contingency tables. The sample maximal correlation that Yenigün formulated is then used to construct a test of independence. Yenigün used exact inferential methods for small sample sizes or contingency tables with sparseness, in which the maximal correlation is applied as the ordering criterion. For the asymptotic null distribution of the test statistic, he used a result by Sethuraman [82].

Another procedure uses the alternating conditional expectations (ACE) algorithm developed by Breiman and Friedman [13]. The ACE algorithm is utilized to find the transformations of Y and X that maximize the proportion of variation in Y explained by X. If X happens to be a matrix, its columns are transformed so that these columns have equal weights when predicting Y. There is an R package called *acepack* developed by Spector, Friedman, Tibshirani and Lumley [89] for its implementation. The `ace` function specifically computes maximal correlation.

CHAPTER 4 THEORETICAL RESULTS

4.1 Comparison of the Properties of Dependence Measures

The following properties outlined by A. Rényi [72] are the properties that most authors inspect or assess in a particular measure of dependence.

A. $\delta(X, Y)$ is defined for any pair of random variables X and Y, neither of them being constant with probability 1.

B. $\delta(X, Y) = \delta(Y, X)$.

C. It should be between 0 and 1; that is, $0 \leq \delta(X, Y) \leq 1$.

D. $\delta(X, Y) = 0$ if and only if X and Y are independent.

E. $\delta(X, Y) = 1$ if there is a strict dependence between X and Y, i.e. either $X = g(Y)$ or $Y = f(X)$, where $g(\cdot)$ and $f(\cdot)$ are Borel-measurable functions.

F. If the Borel-measurable functions $f(\cdot)$ and $g(\cdot)$ map the real axis in a one-to-one way onto itself,

$$\delta(f(X), g(Y)) = \delta(X, Y).$$

G. If the joint distribution of X and Y is normal, then $\delta(X, Y) = |\rho(X, Y)|$ where $\rho(X, Y)$ is the correlation coefficient of X and Y.

Rényi [72] verified these properties on some known measures like Pearson correlation coefficient, correlation ratios, the mean square contingency, and the maximal correlation. He showed that while the first three measures satisfy some of the properties **A**-**G**, only the maximal correlation satisfies all seven of them. Details are included in Section 3.7.

In addition to the measures assessed by Rényi, Linfoot [63] constructed a measure of association based on the amount of information $I(X, Y)$ which X and Y contain with respect to each

Table 4.1: Evaluation of the dependence coefficients in relation to the properties of A. Rényi (✓ means that the property is satisfied, × means that the property is not satisfied, and letter-number code means that the property is partially satisfied and explained further in the text as coded).

	Renyi's Properties	dCor	ρ	$\lvert\rho\rvert$	GGC	RV
A.	$\delta(X,Y)$ is defined for any pair X, Y neither of them being constant with probability 1	A_1	A_2	A_3	A_4	A_5
B.	Symmetric: $\delta(X,Y) = \delta(Y,X)$	✓	✓	✓	✓	✓
C.	$0 \leq \delta(X,Y) \leq 1$	✓	×	✓	✓	✓
D.	$\delta(X,Y) = 0$ if and only if X and Y are independent	✓	D_2	D_3	D_4	D_5
E.	$\delta(X,Y) = 1$ if there is a strict dependence between X and Y	E_1	E_2	E_3	E_4	E_5
F.	$\delta(f(X), g(Y)) = \delta(X,Y)$	F_1	F_2	F_3	F_4	F_5
G.	$\delta(X,Y) = \lvert\rho(X,Y)\rvert$ if the joint distribution of X and Y is normal	G_1	×	✓	✓	G_5

other. The quantity is given by $L(X,Y) = \sqrt{1 - \exp(-2I(X,Y))}$. This quantity satisfies all of the properties that Rényi [72] specified. Some authors such as Schweizer and Wolff [81] and Granger et al. [38] partially modified these properties to suit the parameters they establish.

In this section, we observe and show which of the properties of Rényi are satisfied by the dependence measures chosen in this study. Some of the properties were already proven in the paper where the dependence measure was first introduced, while others are shown here. Table 4.1 gives the summary of the results.

4.1.1 Properties of Distance Correlation $\mathcal{R}$

We verify below which properties of $\mathcal{R}$ satisfy Rényi's properties. Most of the properties are already proven by Székely and Rizzo [92].

A_1 $\mathcal{R}$ partially fulfills axiom **A** because it is defined only for pair of random variables X and Y

with finite first moments, that is, $E|X|_p < \infty$ and $E|Y|_q < \infty$.

B_1 $\mathcal{R}$ satisfies axiom **B** since $\mathcal{R}(\mathbf{X}, \mathbf{Y}) = \mathcal{R}(\mathbf{Y}, \mathbf{X})$.

C_1 $\mathcal{R}$ satisfies axiom **C**; that is, $0 \leq \mathcal{R} \leq 1$.

D_1 $\mathcal{R}$ has property **D** since $\mathcal{R}(\mathbf{X}, \mathbf{Y}) = 0$ if and only if $\mathbf{X}$ and $\mathbf{Y}$ are independent.

E_1 $\mathcal{R}$ partially fulfills axiom **E**. One of the properties of dCor stated by Székely et al. [91, 92] is that if $\mathcal{R}_n(\mathbf{X}, \mathbf{Y}) = 1$, then there exist a vector a, a real number b and an orthogonal matrix C such that $\mathbf{Y} = a + b\mathbf{X}\mathbf{C}$. A counterexample for property **E** is given in the following statements. Let $X \sim N(0, 1)$. Let $Y = X^3$. Then $\mathcal{R}(X, Y) < 1$ because $g(X) = X^3$ is not a linear transformation of X. So $\mathcal{R}(X, Y) < 1$ with $Y = g(X)$.

F_1 $\mathcal{R}$ partially satisfies axiom **F**. Let X, Y be standard normal, with $Y = X$. Then $\mathcal{R}(X, Y) = 1$. Let $f(X) := X$ and $g(Y) := Y^3$. Then both f and g are $1 - 1$ functions that map $\mathbb{R}$ onto itself. In one dimension, $\mathcal{R}(X, Y) = 1$ only if there is a linear relation $aX + bY = c$ between X and Y, for any constants $a, b, c \epsilon \mathbb{R}$, and therefore, $\mathcal{R}(f(X), g(Y)) = \mathcal{R}(X, Y)$ does not hold in general.

G_1 $\mathcal{R}$ partially fulfills axiom **G**. $\mathcal{R}$ has the property **G** only when $\rho = 0$ or $\rho = \pm 1$. In general, if $p = q = 1$ with Gaussian distribution, $\mathcal{R} \leq |\rho|$

$$\mathcal{R}^2 = \frac{\rho \arcsin(\rho) + \sqrt{(1 - \rho^2)} - \rho \arcsin(\frac{\rho}{2}) - \sqrt{(4 - \rho^2)} + 1}{1 + \frac{\pi}{3} - \sqrt{3}} \tag{4.1.1}$$

Hence, this result by Székely et al. [91] is a modification of **G**.

4.1.2 Properties of Maximal Information Coefficient

The maximal information coefficient is a statistic and Reshef et al. [75] do not define a population counterpart. Rényi's properties are used to evaluate population coefficients, not the statistics. Therefore, none of Rényi's properties can be verified.

4.1.3 Properties of Pearson Product-Moment Correlation ρ

Rényi [72] has already shown that ρ satisfies only property $\mathbf{B}$. Here, we include that some of these unsatisfied properties can be made as modifications of Rényi's.

A_2 ρ partially possesses property $\mathbf{A}$. By definition, it is defined for $\sigma_1(x)$ and $\sigma_2(x)$ that are finite and positive.

B_2 ρ fulfills $\mathbf{B}$ by Rényi [72].

C_2 ρ does not fulfill property $\mathbf{C}$ since $-1 \leq \rho \leq 1$.

D_2 ρ only satisfies the sufficient condition of axiom $\mathbf{D}$. It is only if X and Y are independent that $\rho = 0$. The converse is not always true. Following Rényi's example, when $X \sim Uniform(-1, 1)$ and $Y = 5X^3 - 3X$, $\rho(X, Y) = 0$. The following shows that $Cov(X, Y) = 0$, hence, $\rho(X, Y) = 0$:

$$\begin{aligned} Cov(X,Y) &= E(XY) - E(X)E(Y) \\ &= 5E(X^4) - 3E(X^2) - 0(0) \\ &= 5(1/5) - 3(1/3) \\ &= 0 \end{aligned}$$

where

$$E(Y) = 5E(X^3) - 3E(X)$$
$$E(XY) = E(5X^4 - 3X^2) = 5E(X^4) - 3E(X^2)$$

$$
\begin{aligned}
E(X^n) &= \int_{-1}^{1} (1/2) X^n dX \\
&= \frac{1}{2(n+1)} X^{n+1} \Big|_{-1}^{1} \\
&= \begin{cases} \frac{1}{n+1}\ , & n \text{ is even;} \\ 0, & n \text{ is odd.} \end{cases}
\end{aligned}
$$

However, in the case that X and Y are jointly normal, the converse is true.

E_2 ρ partially possesses property **E**. It is true that $\rho = 1$ only if Y is positively linearly related to X. Let $X \sim U(-1, 1)$ and $Y = X^2$. Then $\rho = 0$ because $g(X) = X^2$ is not a linear transformation of X. So $\rho = 0$ with $Y = g(X)$. Hence, property **E** does not hold in general for ρ.

F_2 ρ partially fulfills axiom **F**. Let X, Y be uniformly distributed from (0,1), with $Y = X$. Then $\rho(X, Y) = 1$. Let $f(X) := X$ and $g(Y) := Y^2$. So both f and g are $1 - 1$ functions that map $\mathbb{R}$ onto itself and $\rho(f(X), g(Y)) = 0$. However, $\rho = 1$ only if there is a linear relation $aX + bY = c$ between X and Y, for any constant $b, c \epsilon \mathbb{R}$ and $a \epsilon \mathbb{R}^+$, and therefore $\rho(f(X), g(Y)) = \rho(X, Y)$ is not always true.

G_2 ρ does not satisfy **G**. $\rho \equiv |\rho|$ only if $0 \leq \rho \leq 1$.

4.1.4 Properties of $|\rho|$

Rényi [72] has already shown that $|\rho|$ fulfills only properties **B**, **C** and **G**. Again, we include that some of these unsatisfied properties can be made as modifications of Rényi's.

A_3 $|\rho|$ partially possesses property **A**. By definition, it is defined for $\sigma_1(x)$ and $\sigma_2(x)$ that are finite and positive.

B_3 $|\rho|$ fulfills **B** by Rényi [72].

C_3 $|\rho|$ satisfies **C**.

D_3 $|\rho|$ only satisfies the sufficient condition of axiom $\mathbf{D}$. Same argument as D_2 above where $\rho = 0$, implying that $|\rho| = 0$.

E_3 $|\rho|$ partially possesses property $\mathbf{E}$. Similar to ρ, it is true that $|\rho| = 1$ only if Y is positively linearly related to X. Let $X \sim U(-1, 1)$ and $Y = X^2$. Then $|\rho| = 0$ because $g(X) = X^2$ is not a linear transformation of X. So $|\rho| = 0$ because $\rho = 0$ with $Y = g(X)$. Hence, property $\mathbf{E}$ does not hold in general for $|\rho|$.

F_3 $|\rho|$ partially fulfills axiom $\mathbf{F}$. Same argument as F_2 above. Therefore, $|\rho(f(X), g(Y))| = |\rho(X, Y)|$ is not always true.

G_3 It is obvious that $|\rho|$ possesses $\mathbf{G}$.

4.1.5 Properties of Global Gaussian Correlation

Although the main goal of local Gaussian correlation (LGC) is to estimate the dependence in a neighborhood of a point $\mathbf{x} = (x, y)$, the concern of this paper is to look at how the global correlation coefficient (τ or GGC) would measure the association of the two random variables X and Y. While the LGC satisfies a considerable number of properties such as symmetry, falls between 0 and 1, independence of X and Y implies $\rho_h(X, Y) = 0$ but not vice versa, the properties of the global measure remain to be shown and compared with Rényi's axioms.

A_4 τ partially possesses property $\mathbf{A}$. By definition, τ is defined only when $\sigma_1(x)$ and $\sigma_2(x)$ are finite and positive in each neighborhood of point x.

B_4 τ fulfills $\mathbf{B}$. The distribution of r^2 as formulated using detection theory concept that relates signal to noise ratio (SNR) is shown by Atwood and Spolsky [2] to be noncentral Beta distribution; that is,

$$\begin{aligned} r^2 &\overset{d}{=} \frac{\chi_1^2(\lambda)}{\chi_1^2(\lambda) + \chi_{N_E-1}^2(\lambda^\perp)} \\ &\sim Beta(0.5, 0.5N_E; \lambda, \lambda^\perp) \end{aligned}$$

where the noncentrality parameters $\lambda = \|P_y(x)\|^2/\sigma^2$ and $\lambda^\perp = \|P_y^\perp(x)\|^2/\sigma^2$. When the data stream $\mathbf{x}$ contains only noise, r^2 has a central beta distribution where $\lambda = \lambda^\perp = 0$. In the presence of signal, $\mathbf{x}$ will have a non-zero projection $P_y^\perp(x)$ orthogonal to the noise-contaminated data vector $\mathbf{y}$, hence, $\lambda, \lambda^\perp \neq 0$. Since τ is the sum of r^2 which is beta distributed, then τ is approximately normally distributed, which implies that τ is symmetric.

C_4 τ has property **C** since $0 \leq \tau \leq 1$. From Equation (3.2.3), it is clear that $\tau = E_F(\rho^2(X)) = \int \rho^2(x) dF(x) \geq 0$ for an arbitrary spherical density f, and τ falls between 0 and 1:

$$\begin{aligned}
E_F(\rho^2(X)) &= \int \frac{(\sigma_1(x) - \sigma_2(x))^2}{\sigma_1^2(x) + \sigma_2^2(x) + \sigma_1(x)\sigma_2(x)\left(\tan^2\alpha + \frac{1}{\tan^2\alpha}\right)} dF(x) \\
&\leq \int \frac{(\sigma_1(x) - \sigma_2(x))^2}{\sigma_1^2(x) + \sigma_2^2(x) + 2\sigma_1(x)\sigma_2(x)} dF(x) \\
&= \int \frac{(\sigma_1(x) - \sigma_2(x))^2}{(\sigma_1(x) + \sigma_2(x))^2} dF(x) \\
&\leq \int 1 dF(x) \\
&= 1
\end{aligned}$$

Or alternatively, consider the factor $\sqrt{1 - \rho^2(x)}$ in the joint density (3.2.1). The estimate $\rho_{n,b}(x)$ satisfies $-1 \leq \rho_{n,b}(x) \leq 1$. This implies that $0 \leq \rho_{n,b}^2(x) \leq 1$ and since $\rho_{n,b}(x)$ converges to $\rho(x)$, then $0 \leq \rho^2(x) \leq 1$.

D_4 τ only satisfies the sufficient condition of axiom **D**. Referring to the bivariate normal distribution in Equation (3.2.1), if X and Y are independent, then $f(x, y) = f(x)f(y)$ implying that $f(x, y) - f(x)f(y) = 0$. Since

$$\begin{aligned}
f(x) &= \frac{1}{\sqrt{2\pi}\sigma_1(x)} \exp\left(-\frac{1}{2\sigma_1^2(x)}(x - \mu_1(x))^2\right) \\
f(y) &= \frac{1}{\sqrt{2\pi}\sigma_2(y)} \exp\left(-\frac{1}{2\sigma_2^2(y)}(y - \mu_2(y))^2\right)
\end{aligned}$$

which means that in the neighborhood of each point (x, y),

$$\begin{aligned}\sqrt{1-\rho^2(x,y)} = 1 &\Rightarrow 1-\rho^2(x,y) = 1 \\ &\Rightarrow \rho^2(x,y) = 0 \\ &\Rightarrow \tau(x,y) = E[\rho^2(x,y)] = 0.\end{aligned}$$

E_4 τ partially possesses property **E**. Tjøstheim and Hufthammer [99] specified that if X and Y are linearly related not necessarily Gaussian with $Y = a+bX$, then $\rho = 1$ if $b > 0$ and $\rho = -1$ if $b < 0$. However, for $Y = h(X)$ for some function h, the value of ρ depends on h. Tjøstheim and Hufthammer gave an example that for $Y = X^2$, $\rho = 0$ under weak assumptions. Also, E_2 and E_3 show counterexamples of how τ violates this axiom in general.

F_4 τ partially fulfills axiom **F** when f and g are linear functions. Let $Y = X$. Then, $\rho(X, Y) = 1$. Using similar example in D_2, let $X \sim Uniform(-1, 1)$, we get $Var(X) = 1/3$ and $Var(X^3) = 1/7$. So if $f(X) = X$, $g(Y) = Y^3$ and $X = Y$,

$$\begin{aligned}\rho(f(X), g(Y)) &= \rho(X, X^3) \\ &= \frac{E(X^4) - E(X)E(X^3)}{\sqrt{Var(X)Var(X^3)}} \\ &= \frac{1/5 - 0}{\sqrt{(1/3)(1/7)}} \\ &= \frac{1}{5\sqrt{21}} \\ &< 1.\end{aligned}$$

But, $\rho(X, Y) = 1$, which means that $\rho(f(X), g(Y)) \neq \rho(X, Y)$ in this case. Thus, **F** is not true for ρ and not true for $|\rho|$ as well. Therefore, τ does not have property **F** in general.

G_4 τ satisfies **G**. Berentsen et al. [6] specified this particular property: If the joint distribution of X and Y is Gaussian with constant correlation coefficient ρ, then $\tau \equiv |\rho|$.

4.1.6 Properties of the RV coefficient

We verify below which properties of ρV coefficient satisfy Rényi's properties.

A_5 ρV partially satisfies axiom **A**. From the definition of ρV in (3.4.1), it is clear that ρV is defined for any pair of random variables X and Y when $\Sigma_{XX}^2 \Sigma_{YY}^2 \neq 0$.

B_5 ρV fulfills axiom **B**, that is, $\rho V(X, Y) = \rho V(Y, X)$. It can be easily shown utilizing the properties of trace that the definition of ρV in Equation (3.4.1) can be written as

$$\begin{aligned}\rho V(X, Y) &= \frac{tr(XX'YY')}{\sqrt{tr(XX')^2 tr(YY')^2}} \\ &= \frac{tr(YY'XX')}{\sqrt{tr(YY')^2 tr(XX')^2}} \\ &= \rho V(Y, X).\end{aligned}$$

C_5 ρV satisfies axiom **C**. Abdi [1] has proved this property.

D_5 ρV partially possesses axiom **D**. If X and Y are independent, then $\Sigma_{XY} = 0$. Therefore, $\rho V = 0$. However, as mentioned by Josse and Holmes [51], $\rho V = 0$ does not necessarily imply that X and Y are independent unless an assumption of multivariate normality is satisfied.

E_5 ρV partially satisfies axiom **E** when X, Y are assumed normal and $Y = g(X) = aBX + c$ where a is a constant, B is an orthogonal matrix and c is a constant vector. Using the properties of covariance and variance,we show below that $\rho V(X, Y) = 1$. This is true since Y is linear as defined above. When X and Y have multivariate normal distributions but

$Y = g(X)$ is not linear, $\rho V(X,Y) < 1$.

$$\begin{aligned}\rho V(X,Y) &= \rho V(X, aBX + c) \\ &= \frac{CovV(X, aBX + c)}{\sqrt{VaV(X)VaV(aBX + c)}} \\ &= \frac{aBCovV(X,X)}{\sqrt{VaV(X)aBVaV(X)B^T a^T}} \\ &= \frac{aBVaV(X)}{aBVaV(X)} \\ &= 1.\end{aligned}$$

F_5 ρV partially fulfills axiom **F**. Let X, Y be standard normal, with $Y = X$. Then $\rho V(X,Y) = 1$. Let $f(X) := X$ and $g(Y) := Y^3$. Then, both f and g are $1-1$ functions that map $\mathbb{R}$ onto itself. By the construction of ρV, considering one dimension, it can only detect dependence only if there is a linear relation $aX + bY = c$ between X and Y, for any constants $a, b, c \epsilon \mathbb{R}$, which means that $\rho V < 1$. Therefore, in general, $\rho V(f(X), g(Y)) = \rho V(X,Y)$ is not true.

G_5 ρV partially satisfies property **G** because $\rho V \neq |\rho|$ but ρV is a function of ρ. When $p = q = 1$ and X, Y are normally distributed, $\rho V = \rho^2$ where ρ^2 is the square of the simple correlation coefficient. Equation (3.4.1) can be rewritten as

$$\begin{aligned}\rho V(X,Y) &\frac{tr(\sigma_{XY}\sigma_{YX})}{\sqrt{tr(\sigma_{XX}^2)tr(\sigma_{YY}^2)}} \\ &= \frac{\sigma_{XY}\sigma_{YX}}{\sqrt{\sigma_{XX}^2\sigma_{YY}^2}} \\ &= \rho^2.\end{aligned}$$

The dependence measures that we have considered possess few or some of the seven properties that Rényi formulated as an axiomatic framework. However, the most important property that a particular dependence measure should fulfill is axiom **D**, which states that, $\delta(X,Y) = 0$ if and only if X and Y are independent. It is intuitive that the coefficient should give a value of 0 to

Table 4.2: Evaluation of the dependence coefficients in terms of desirable properties

Properties	dCor	ρ	GGC	MIC	RV
Rigid Motion Invariance	✓	×	✓	×	✓
R^2-Equitability	×	×	×	×	×
Self-Equitability	×	×	×	×	×
DPI	×	×	×	×	×
Scale Invariance	✓	✓	✓	×	✓
Bivariate Measure	✓	✓	✓	✓	×
Multivariate Measure	✓	×	×	×	✓
High-Dimensional	✓	×	×	✓	×
Consistent	✓	✓	✓	–	✓

any random vectors X and Y that are independent, and on the other hand, if the coefficient is 0, it should imply independence. Otherwise, the dependence measure does not convey a meaningful information. The global Gaussian correlation, RV coefficient, and Pearson partially satisfy the sufficient condition of $\mathbf{D}$. The only measure that fully satisfies $\mathbf{D}$ is the distance correlation.

4.2 Desirable Properties of Dependence Measures

Table 4.2 summarizes other properties that are desirable for measures of dependence. It shows which of the measures considered and studied here possess the said properties.

4.2.1 Property of Equitability

It was claimed by Reshef et al. [75, 76] that the maximal information coefficient (MIC) they developed possesses a desirable mathematical property called "equitability".

Reshef et al. stated that a good measure of dependence should possess the property of equitability; that it should give similar scores to equally noisy relationships regardless of relationship types. This means that there is independence between the measure of how much noise is in an x-y scatter plot and on what the specific functional relationship between x and y would be in the absence of noise. They added that equitability is difficult to formalize for associations in general but has a clear interpretation in the basic case of functional relationships. That is, given sufficient

sample size n, an equitable statistic should give similar scores to functional relationships with similar values of sample coefficient of determination, R^2. Reshef et al. [76] considered a setting that corresponds to sampling in which both coordinates are subject to noisy requirements. Specifically, the data take the form $(X + N_x, f(X) + N_y)$ where X are uniformly distributed over $[0, 1]$, N_x and N_y are uniformly distributed in a small interval and independent of each other and of X. In this scenario, a measure of dependence δ is equitable to the extent that the R^2 of $(X + N_x, f(X) + N_y)$ with respect to the function f depends only on the score assigned by δ to $(X + N_x, f(X) + N_y)$ (not on f), and vice versa.

According to Reshef et al. [76], equitability is important in exploration of high dimensional data sets, where there can be many pairwise relationships to capture and there is no valid reason to favor certain types of relationships over others. However, a host of questions and criticisms arose, so that some researchers made an in-depth study and analysis of this concept. Kinney and Atwal [54] made a thorough investigation of MIC and its properties, specifically the property of equitability.

Kinney and Atwal [54] called it R^2-equitability and formally defined it as follows:

Definition 4.1. *A dependence measure, $D[X;Y]$, is R^2-equitable if and only if, when evaluated on a joint probability distribution $p(X, Y)$ that corresponds to a noisy functional relationship between two real random variables X and Y, the following relation holds:*

$$D[X;Y] = g(R^2[f(X);Y]), \tag{4.2.1}$$

whenever

$$Y = f(X) + \eta. \tag{4.2.2}$$

Here, g is an unspecified function that does not depend on f, f is a deterministic function of X and η is a random noise term. The noise term η may depend on $f(X)$ as long as η has no additional dependence on X, i.e., as long as $X \leftrightarrow f(X) \leftrightarrow \eta$ is a Markov chain.

Careful examination of the definition of equitability which Kinney and Atwal [54] called as R^2-equitability led to the following conclusion: *There is no nontrivial dependence measure that can satisfy R^2-equitability.* The reason is due to the fact that the function defining the noisy functional relationship in 4.2.2 is not uniquely specified by a joint probability distribution $p(X, Y)$. They considered an example on a simple functional linear relationship $Y = X + \eta$. For every invertible function f, there also exists a valid noise term ξ such that $Y = f(X) + \xi$. Thus, R^2-equitability requires that a dependence measure satisfies $D[X;Y] = g(R^2[X;Y]) = g(R^2[f(X);Y])$. But, $R^2[X;Y]$ is not invariant under invertible transformations of X. Hence, the function g must be constant, implying that $D[X;Y]$ does not depend on $p(X, Y)$ and is therefore trivial. Therefore, dCor, GGC, Pearson's r, RV and MIC are not R^2-equitable.

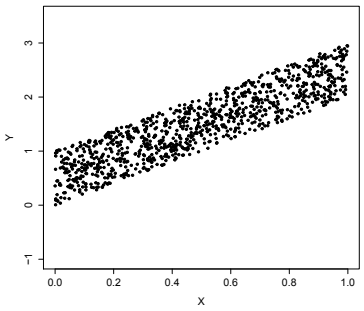
Table 4.3 illustrates the idea of equitability. It includes the plots of four different functional dependence where 1000 data points are simulated with equal noise. The relationships considered are linear of the form $Y = 2X' + 0.5 + \epsilon'$ where $X' \sim U(0, 1)$, quadratic of the form $Y = X^2 + 0.5 + \epsilon$, cubic of the form $Y = X^3 + 0.5 + \epsilon$ and exponential of the form $Y = e^X + 0.5 + \epsilon$ where $X \sim U(-1, 1)$ and $\epsilon', \epsilon \sim U(-0.5, 0.5)$. Table 4.3 shows the corresponding $mean \pm sd$ values of the four measures computed from 1000 replicates. When a dependence measure satisfies the property of equitability, it assigns similar scores to these relationships. However, there is not one measure which assigns similar scores to all four relationships. These results confirmed the conclusions made by Kinney and Atwal [54], that there is no nontrivial dependence measure can satisfy R^2-equitability.

Since R^2-equitability was not found useful and adaptable, Kinney and Atwal formalize the notion of equitability as an invariance property and they called it self-equitability which is defined in Definition 4.2.

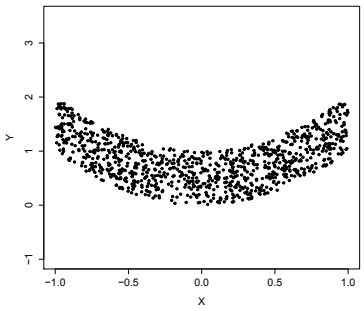
According to Kinney and Atwal, the definitions of self-equitability and R^2-equitability are somewhat similar except for three points. First, the noise in the relationship is quantified using D itself rather than R^2. Second, Y can be of any type like categorical, multidimensional or non-commutative. Third, Y depends on X only through the value of $f(X)$ not on the additive noise η.

Table 4.3: Illustration of four different dependence structures with equal noise including a table that gives the mean and standard deviation of the four dependence measures.

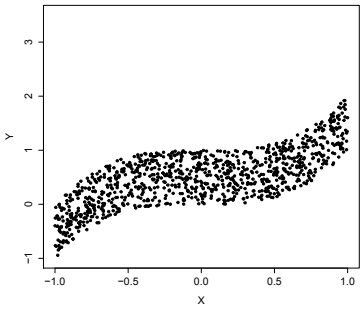
(a) Linear



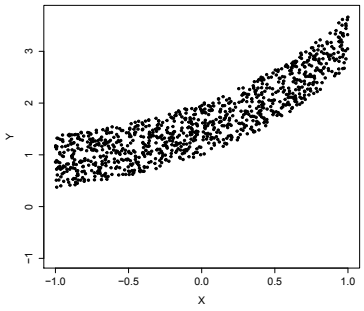
(b) Quadratic



(c) Cubic



(d) Exponential



Measure	Linear	Quadratic	Cubic	Exponential
dCor ($\mathcal{R}$)	0.79 ± 0.03	0.15 ± 0.03	0.43 ± 0.07	0.78 ± 0.04
GGC (τ)	0.90 ± 0.01	0.27 ± 0.03	0.73 ± 0.05	0.90 ± 0.02
MIC	0.82 ± 0.06	0.47 ± 0.08	0.51 ± 0.06	0.82 ± 0.06
Pearson (r)	0.80 ± 0.03	0.01 ± 0.02	0.53 ± 0.06	0.79 ± 0.03

These advantages of self-equitability result to the fact that any self-equitable dependence measure is invariant under arbitrary invertible transformations of X or Y.

Definition 4.2. *A dependence measure $D[X;Y]$ is self-equitable if and only if it is symmetric, i.e. $D[X;Y] = D[Y;X]$, and satisfies*

$$D[X;Y] = D[f(X);Y], \tag{4.2.3}$$

whenever f is a deterministic function, X and Y are variables of any type, and $X \leftrightarrow f(X) \leftrightarrow Y$ forms a Markov chain.

Self-equitability is closely related to the Data Processing Inequality (DPI), a very important feature of information theory. DPI formalizes Kinney and Atwal's idea that when information is transferred through a noisy communications channel, information is lost and not anymore gained. A dependence measure that is DPI is defined as follows.

Definition 4.3. *A dependence measure $D[X;Y]$ satisfies DPI if and only if:*

$$D[X;Y] \leq D[f(X);Y], \tag{4.2.4}$$

whenever the random variables X, $f(X)$, Y form a Markov chain $X \leftrightarrow f(X) \leftrightarrow Y$.

It is shown that every dependence measure $I_F[X;Y]$ that can be written as

$$I_F[X;Y] = \int p(x)p(y)F\left(\frac{p(x,y)}{p(x)p(y)}\right)dxdy, \tag{4.2.5}$$

where F is a convex function on the nonnegative real numbers, satisfies DPI. All dependence measures that satisfy DPI are also self-equitable. But there are self-equitable measures that do not satisfy DPI. Though these measures exist, it is still reasonable to require that measures satisfy DPI because DPI implements a significant heuristic that self-equitability does not. Kinney and Atwal [54] stated that mutual information satisfies DPI but MIC does not.

We will show below that RV coefficient does not satisfy self-equitability and DPI. Without loss of generality, assume $p = q = 1$. Let $X \leftrightarrow f(X) = X^2 \leftrightarrow Y$ form a Markov Chain. As shown in Section 4.1.6, RV satisfies symmetry. However, if $XX' \neq X^2(X^2)'$,

$$\begin{aligned} RV(X,Y) &= \frac{tr(XX'YY')}{\sqrt{tr(XX')^2 tr(YY')^2}} \\ &\neq \frac{tr(X^2(X^2)'YY')}{\sqrt{tr(XX')^2 tr(YY')^2}} \\ &= \frac{tr(X^2(X^2)'YY')}{\sqrt{tr(XX')^2 tr(YY')^2}} \cdot \frac{\sqrt{tr(X^2(X^2)')^2}}{\sqrt{tr(X^2(X^2)')^2}} \\ &= \frac{tr(X^2(X^2)'YY')}{\sqrt{tr(X^2(X^2)')^2 tr(YY')^2}} \cdot \frac{\sqrt{tr(X^2(X^2)')^2}}{\sqrt{tr(XX')^2}} \\ &= RV(X^2,Y) \cdot k \end{aligned}$$

for any constant $k > 0$. This implies that $RV(X,Y) \neq RV(X^2,Y)$. Thus, RV coefficient does not fulfill self-equitability. Moreover, $RV(X,Y) \not\leq RV(X^2,Y)$ for $0 < k < 1$. Therefore, RV does not satisfy DPI.

4.2.2 Property of Rigid Motion Invariance

The property of rigid motion invariance is another property that is desirable for the dependence measures. Rigid motion invariance also refers to location and rotation invariance. It is intuitive that the measure of dependence of random variables X and Y should not change even when the units of measurement of X and Y are changed or when the orthogonal bases are changed.

Definition 4.4. *Let $x, y \epsilon \mathbb{R}^p$ and $T : \mathbb{R}^p \to \mathbb{R}^p$ be a rigid transformation defined by $T(u) = Au + b$ where A is a linear, orthogonal matrix and $b \epsilon \mathbb{R}^p$. Then a dependence measure $D[X;Y]$ is rigid motion invariant if the value of $D[X;Y]$ remains constant under rigid transformation.*

Remark 4.5. *1. A dependence measure $D[X;Y]$ that preserves angles and lengths of vectors is rigid motion invariant.*

2. If X and Y in $\mathbb{R}^p$ are both nonzero, then the angle between X and Y, denoted by $\angle(X,Y)$,

is defined to be $\arccos(\langle x, y\rangle/|x||y|)$. *Angle is preserved if the following holds for any rigid transformation* T*:*

$$\begin{aligned}\frac{\langle T(x), T(y)\rangle}{|T(x)||T(y)|} &= \frac{\langle T(x) - T(0), T(y) - T(0)\rangle}{|T(x) - T(0)||T(y) - T(0)|} \\ &= \frac{\langle Ax, Ay\rangle}{|Ax||Ay|} \\ &= \frac{\langle x, y\rangle}{|x||y|}. \end{aligned} \tag{4.2.6}$$

On the other hand, distance or length, d*, is preserved if:*

$$d(T(x), T(y)) = d(x, y). \tag{4.2.7}$$

3. *Euclidean norm is preserved under rigid motion since it has the following property:*

$$|T(u)| = |u|, \tag{4.2.8}$$

where $u = (x, y)$. *We will see that distance-based measures dCor and RV are rigid motion invariant.*

4. *In general, the structures of Euclidean space which includes distance, angles, lines and vectors are invariant under the transformations of their associated Euclidean group. The Euclidean group treats rotations, translations, reflections in a similar way.*

The next paragraphs show the dependence measures that are invariant and not invariant to rigid motion.

We have seen in Chapter 3 that distance correlation (dCor) is a distance-based measure. Distance covariance as defined in (3.1.1) is constructed based on a weighted L^2 norm and the statistic is a function of distances $d_{i,j}$. Since the dCov statistic given by

$$\mathcal{V}_n^2 = S_1 + S_2 - 2S_3,$$

where

$$S_1 = \frac{1}{n^2}\sum_{k,l=1}^{n} |X_k - X_l|_p |Y_k - Y_l|_q,$$

$$S_2 = \frac{1}{n^2}\sum_{k,l=1}^{n} |X_k - X_l|_p \frac{1}{n^2}\sum_{k,l=1}^{n} |Y_k - Y_l|_q,$$

$$S_3 = \frac{1}{n^3}\sum_{k=1}^{n}\sum_{l,m=1}^{n} |X_k - X_l|_p |Y_k - Y_m|_q,$$

are linear combination of distances, and dCor is the standardized dCov, then it follows that dCor is rigid motion invariant.

As discussed in Section 3.4, Robert and Escoufier [79] proposed that RV coefficient can be viewed as a unifying tool to analyze some of the classical linear multivariate statistical methods. Robert and Escoufier derived this idea as the quest for optimal linear transformations or, equivalently, the quest for optimal metrics to apply on two data matrices on the same sample. The optimality is defined in terms of the similarity of the corresponding configurations of points, which implies the maximization of the associated RV-coefficient.

Robert and Escoufier showed that the distance between two configurations $C(X)$ and $C(Y)$ of a $p \times n$ data matrix X and a $q \times n$ data matrix Y, respectively, can be written as a function of RV.

$$\begin{aligned} dist\{C(X), C(Y)\} &= \|S(X)/\{trS(X)^2\}^{1/2} - S(Y)/\{trS(Y)^2\}^{1/2}\| \\ &= \sqrt{2\left(1 - \frac{tr\{S(X)S(Y)\}}{trS(X)^2 trS(Y)^2}\right)} \\ &= \sqrt{2\{1 - RV(X,Y)\}} \end{aligned}$$

The matrix $S(X)/\{trS(X)^2\}^{1/2}$ was preferred because it is translation and rotation independent. The same is true with $S(Y)/\{trS(Y)^2\}^{1/2}$. Then, the distance between the two configurations $dist\{C(X), C(Y)\}$ is also translation and rotation independent. This implies that RV is also trans-

lation and rotation independent, or equivalently, RV is rigid motion invariant.

By looking at the definition of Pearson product-moment correlation and the properties of covariance and variance, we conclude that Pearson product-moment correlation is not invariant to rigid motion. First, ρ is translation invariant; that is,

$$\rho(a + X, b + Y) = \rho(X, Y), \tag{4.2.9}$$

because $Cov(a + X, b + Y) = Cov(X, Y)$, $Var(a + X) = Var(X)$ and $Var(b + Y) = Var(Y)$. Second, ρ is scale invariant; that is,

$$\rho(aX, bY) = \frac{abCov(X, Y)}{\sqrt{a^2 Var(X) \cdot b^2 Var(Y)}} = \rho(X, Y), \tag{4.2.10}$$

since $Cov(aX, bY) = abCov(X, Y)$, $Var(aX) = a^2 Var(X)$ and $Var(bY) = b^2 Var(Y)$. However, $Cov(-X, Y) = -Cov(X, Y)$ so,

$$\rho(-X, Y) \neq \rho(X, Y), \tag{4.2.11}$$

which implies that invariance to reflection fails. Thus, ρ is not rigid motion invariant.

The global Gaussian correlation coefficient τ defined in (3.2.4) is rigid motion invariant. Since τ is the sum of all ρ^2 computed in each neighborhood of x, it is important to know if ρ^2 is rigid motion invariant. Based on Equations (4.2.9) and (4.2.10), it is true that ρ^2 is location and scale invariant; that is,

$$\rho^2(a + X, b + Y) = \rho^2(X, Y), \tag{4.2.12}$$

$$\rho^2(aX, bY) = \rho^2(X, Y), \tag{4.2.13}$$

respectively. In addition, the ρ^2 coefficient fulfills invariance to reflection because

$$\begin{aligned}\rho^2(-X,Y) &= \left(\frac{Cov(-X,Y)}{\sqrt{Var(-X)\cdot Var(Y)}}\right)^2 \\ &= \left(\frac{-Cov(X,Y)}{\sqrt{Var(X)\cdot Var(Y)}}\right)^2 \\ &= \left(\frac{Cov(X,Y)^2}{Var(X)\cdot Var(Y)}\right) \\ &= \rho(X,Y)^2. \end{aligned} \tag{4.2.14}$$

Similarly, $\rho^2(X,-Y) = \rho(X,Y)^2$ and $\rho^2(-X,-Y) = \rho(X,Y)^2$. Therefore, τ is rigid motion invariant.

An alternative proof that we derive from what Tjøstheim et al. [99] wrote in their paper is the following. An arbitrary spherical density f is considered for a given vector x as well as the rotation matrix

$$A = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix}.$$

Then, Ax is rotated counter-clockwise through an angle α. This f is a rotation symmetric in addition to being radial, reflection and exchange symmetric. Tjøstheim et al. defined

$$\begin{aligned}\rho^2 &= \rho^2(\alpha) \\ &= \frac{\{\sigma_{11}(x)-\sigma_{22}(x)\}^2}{\sigma_{11}^2(x)+\sigma_{22}^2(x)+\sigma_{11}(x)\sigma_{22}(x)\left(\tan^2(\alpha)+\frac{1}{\tan^2(\alpha)}\right)}\end{aligned}$$

which has its maximum for $\tan^2(\alpha) = 1$; that is, $\alpha = \pm\pi/4$. Note that $\rho^2(\alpha)$ is positive in all quadrants. The region of f is conformally mapped by a Riemann mapping onto the unit disk implying that the angles are preserved and hence, values of ρ^2 remain unchanged under rigid motion. Thus, the sum of all ρ^2, which is τ, is invariant to rigid motion.

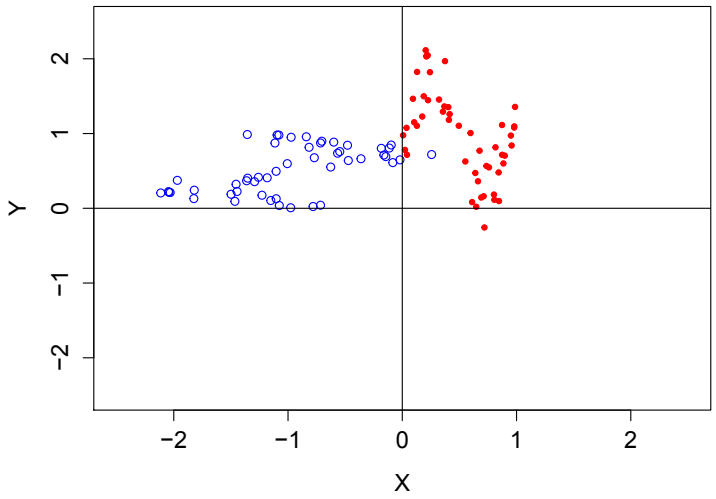
In contrast, MIC defined in (3.3.1) is not rigid motion invariant. Reshef et al. [75] discussed in their supplementary online material that MIC is not invariant under rotation of the coordinate

axes. They provided an example where the plot of a slightly noisy diagonal line exhibits statistical dependence, however, if the diagonal line is rotated so that it is horizontal, the plot will exhibit statistical independence. This is also similar given a sufficient sample size, the former plot will have a non-zero MIC while the latter plot will have an MIC that is very close to 0. In addition, we presume that since MIC is a method that uses a binning scheme, the discretization of the data are affected when the data points are rotated, thus altering the value of MIC. We show an example with simulated data in Appendix A.

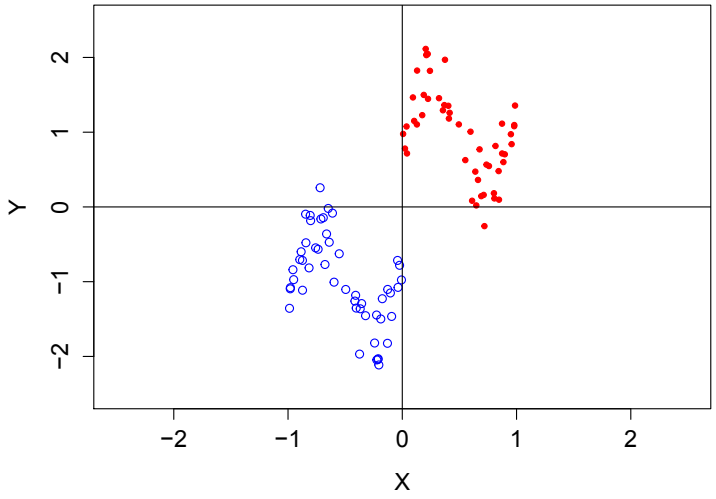
Figure 4.1 demonstrates rotation, a type of rigid motion, specifically 90° and 180°. Figure 4.2 demonstrates reflection, translation and scale transformation. The figures consider the sinusoidal relationship of X and Y, where X has a standard uniform distribution and $Y = \sin(2\pi X) + X + \epsilon$ in which ϵ follows also a standard uniform distribution.

The coefficient of a good measure of dependence should remain the same no matter what transformation (reflection, translation and scale) is made to the values or data points.

Figure 4.1: Illustration of Rigid Motion (Rotations) of $X \sim U(0,1)$ and $Y = \sin(2\pi X) + X + \varepsilon$ where $\varepsilon \sim U(0,1)$ drawn in dots. Figure A. 90° rotation is drawn in circles. Figure B. 180° rotation is drawn in circles.

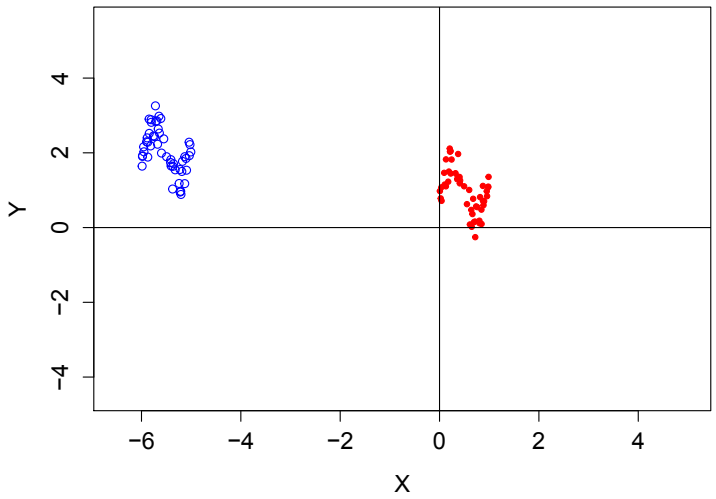


(a)

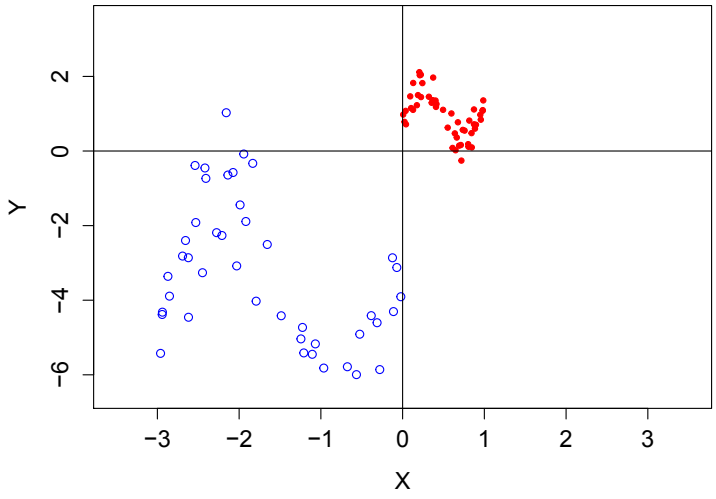


(b)

Figure 4.2: Illustration of Rigid Motion (Translation and Scale) of $X \sim U(0,1)$ and $Y = \sin(2\pi X) + X + \varepsilon$ where $\varepsilon \sim U(0,1)$ drawn in dots. Figure A. Translation $(-X-5, -Y+3)$ is drawn in circles. Figure B. Reflection and scale transformations $(-3X, -4Y)$ is drawn in circles.



(a)



(b)

CHAPTER 5 EMPIRICAL RESULTS

5.1 Comparison of Dependence Measures

In this section we compare how the four statistics - distance correlation (dCor), global Gaussian correlation (GGC), maximal information coefficient (MIC) and Pearson product moment correlation (r) - describe different dependence structures and discuss how some of their properties support the results. The dependence structures considered are displayed in Figure 5.1. Each dependence measure is computed from 1000 replicates for a sample of size $n = 500$. A violin plot, which shows a box plot with a rotated kernel density, is utilized to display the comparison of the measures for each dependence structure. Table 5.1 displays the summary statistics of the coefficients of dCor, GGC, MIC and Pearson. This provides the mean, standard deviation and coefficient of variation of the four measures of dependence.

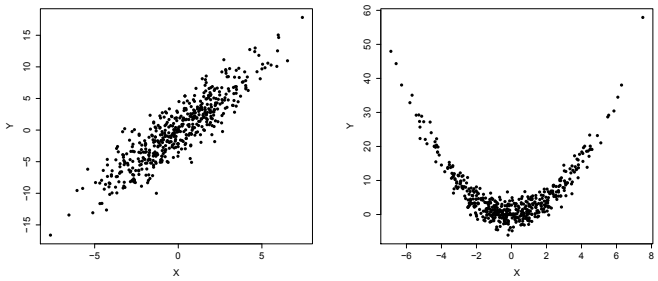
5.1.1 Linear dependence

In this example, the distribution of X is normal with mean 0 and standard deviation 2.5, and $Y_i = 2X_i + \varepsilon_i$, where ε_i are independent normal variables with mean 0 and standard deviation 2.5 and independent of X. Figure 5.1a exhibits the linear plot. The violin plot shown in Figure 5.2 describes the distribution of the four methods along with the summary of the estimates when $n = 500$. It can be seen that the Pearson product moment correlation r is superior in estimating linearity as its definition indicates. The second method that does a good job in characterizing linearity is the dCor followed by GGC. This suggests that dCor and GGC perform well in detecting a linear relationship. However, the distribution of GGC replicates is wider than the distribution of dCor replicates. Maximal information coefficient (MIC) describes linearity the least well among the four, with high variability. As MIC ranks last in capturing simple linear dependence, the result suggests that MIC does not serve its purpose as a tool for exploratory data analysis, which is suppose to detect as many pairwise relationships as possible.

Table 5.1: Summary statistics table of the dependence coefficients measuring different models of X and Y of sample size 500 with 1000 replicates

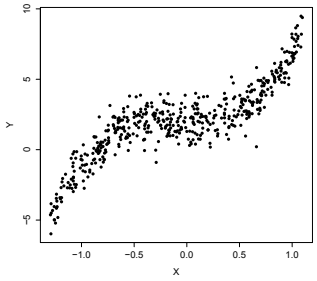
Model	Dependence Measure	Mean (M)	Standard Deviation (SD)	Coefficient of Variation (CV)
Linear	dCor	0.861	0.012	0.014
	GGC	0.847	0.020	0.024
	MIC	0.688	0.034	0.050
	Pearson	0.894	0.009	0.010
Quadratic	dCor	0.494	0.013	0.026
	GGC	0.851	0.031	0.036
	MIC	0.669	0.032	0.048
	Pearson	-0.002	0.096	- 40.125
Cubic	dCor	0.821	0.015	0.018
	GGC	0.868	0.014	0.016
	MIC	0.614	0.027	0.044
	Pearson	0.850	0.011	0.013
Exponential	dCor	0.860	0.012	0.014
	GGC	0.875	0.015	0.018
	MIC	0.686	0.033	0.048
	Pearson	0.868	0.011	0.012
Sinusoidal	dCor	0.398	0.025	0.063
	GGC	0.383	0.039	0.101
	MIC	0.872	0.026	0.030
	Pearson	- 0.360	0.036	- 0.101
Diamond	dCor	0.143	0.013	0.092
	GGC	0.095	0.011	0.110
	MIC	0.172	0.013	0.075
	Pearson	- 0.004	0.029	- 73.000
Four Independent Clouds	dCor	0.015	0.027	1.753
	GGC	0.211	0.025	0.117
	MIC	0.163	0.012	0.075
	Pearson	0.00004	0.044	1100.00
Independent-*t*	dCor	0.041	0.024	0.590
	GGC	0.347	0.052	0.150
	MIC	0.156	0.013	0.080
	Pearson	0.073	0.044	0.595

Figure 5.1: Different bivariate dependence structures considered for comparison

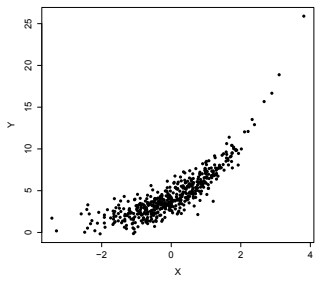


(a) Linear

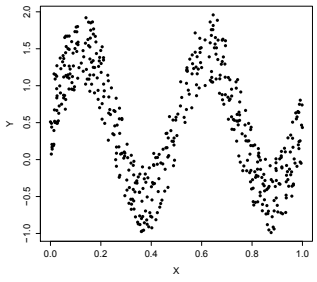
(b) Quadratic



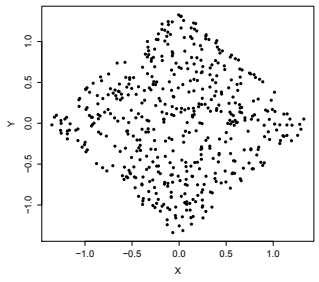
(c) Cubic



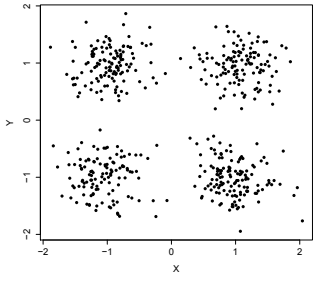
(d) Exponential



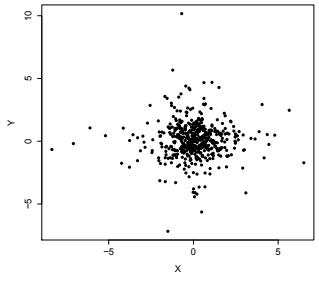
(e) Sinusoid



(f) Diamond



(g) Four Independent Clouds



(h) Independent t

Figure 5.2: Comparison of the measures describing linear dependence: $Y = 2X + \varepsilon$, where $X, \varepsilon \sim N(0, 2.5)$ are independent, using sample size of 500 with 1000 replicates

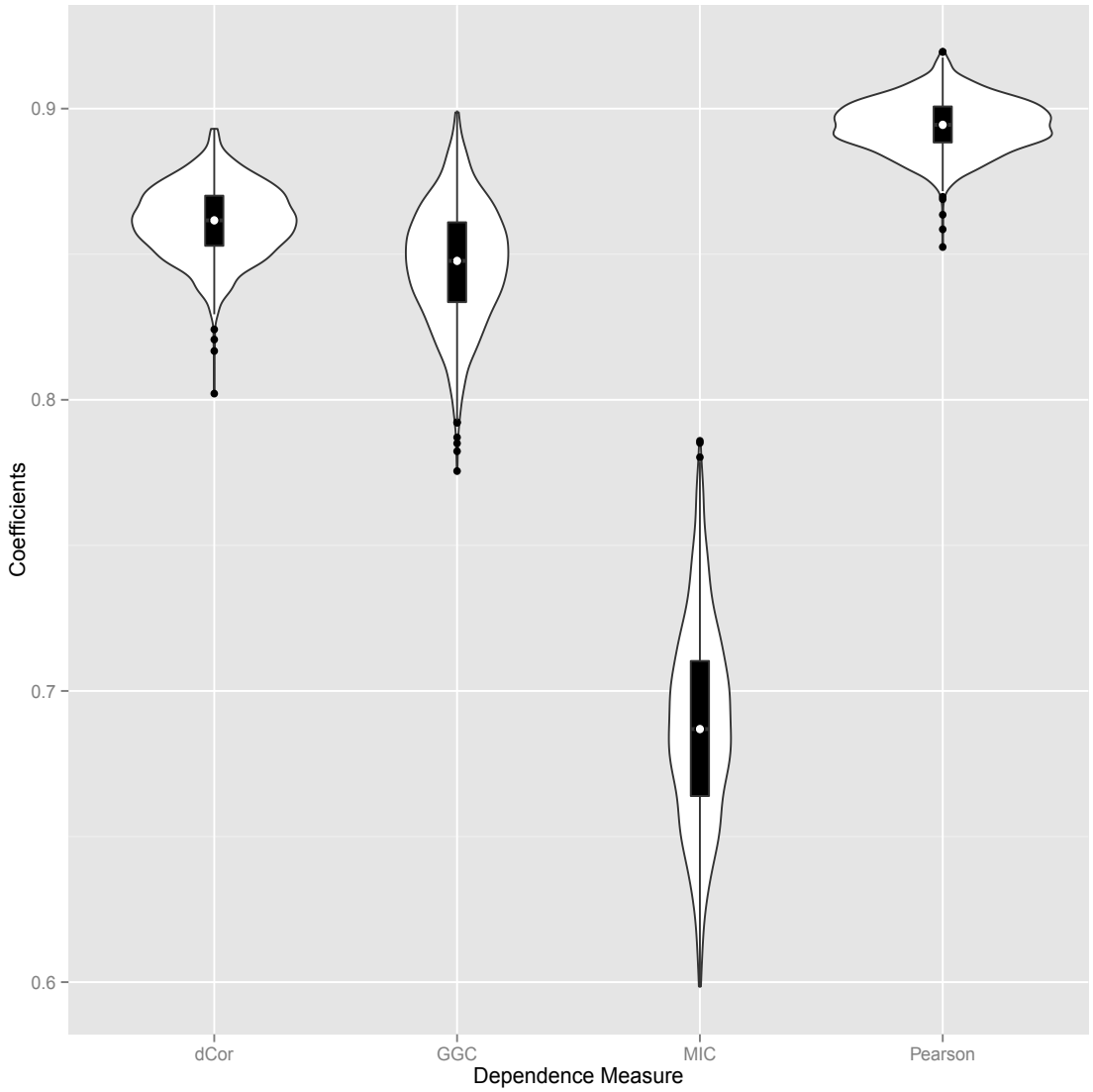
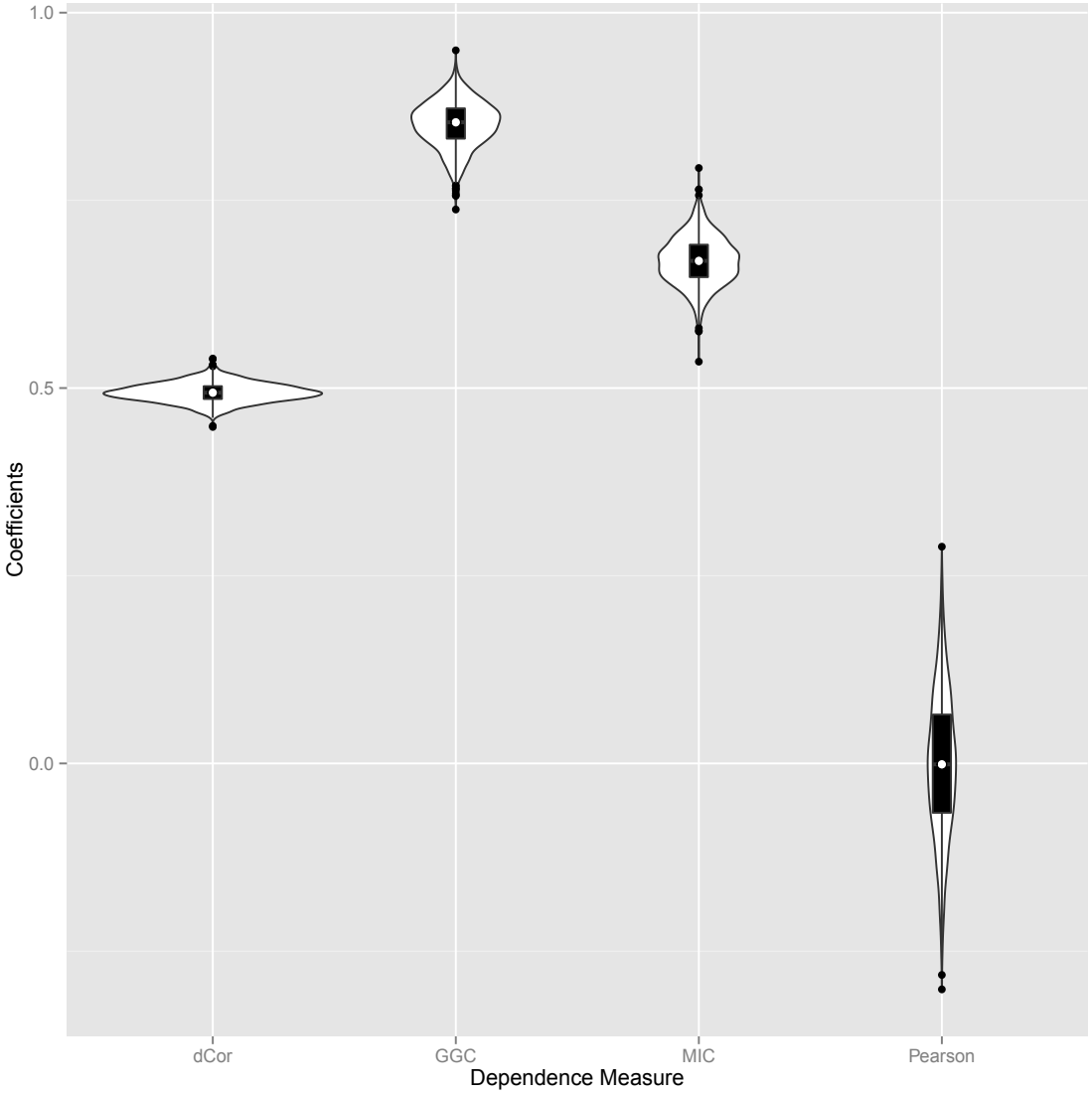


Figure 5.3: Comparison of the measures describing quadratic dependence: $Y = X^2 + \varepsilon$, where $X, \varepsilon \sim N(0, 1.5)$ are independent, using sample size of 500 with 1000 replicates



5.1.2 Quadratic dependence

Here, the distribution of X is normal with mean 0 and standard deviation 1.5, and $Y_i = X_i^2 + \varepsilon_i$, where ε_i are independent normal variables with mean 0 and standard deviation 1.5 and independent of X. Figure 5.1b displays the quadratic plot. The violin plot in Figure 5.3 shows that the estimates of GGC exhibit values close on the average to 0.85, detecting the relationship accurately. MIC and dCor follow with an average of 0.67 and 0.5, respectively. Pearson r fails to detect the association in this example.

5.1.3 Cubic Model

In this example, the distribution of X is uniformly distributed from $(-1.3, 1)$ and $Y_i = 4X_i^3 + X_i^2 + \varepsilon_i$, where ε_i are independent normal variables with mean 1.5 and standard deviation 0.85 and independent of X. The cubic model is plotted in Figure 5.1c. The violin plot in Figure 5.4 indicates that the estimates of GGC, dCor and Pearson do not differ much in measuring the cubic relationship, although Pearson's r surprisingly captures the relationship. This might be due to the fact that cubic function is monotonic like linearity. MIC, on the other hand, ranks last with the distribution of its estimates having wider spread.

5.1.4 Exponential Model

The exponential model in Figure 5.1d is simulated as follows. The distribution of X is standard normal and $Y_i = 4\exp(0.5X_i) + 2 + \varepsilon_i$, where ε_i are independent standard normal variables and independent of X. The violin plot presented in Figure 5.5 displays that the three coefficients, dCor, GGC and Pearson's r, each detected the exponential association between X and Y fairly well in almost the same manner since their estimates are close to 1. The mean estimate of MIC is below 0.8 with the largest spread compared to the other three. This shows that using MIC cannot guarantee a good estimate of this type of dependence. It is interesting to note, however, that although Pearson's r measures only linear dependence, it is able to detect with higher coefficients than MIC for exponential dependence, which is mainly due to the fact that exponential function is

Figure 5.4: Comparison of the measures describing cubic model: $Y_i = 4X_i^3 + X_i^2 + \varepsilon_i$, where $X \sim U(-1.3, 1)$, $\varepsilon \sim N(1.5, 0.85)$ are independent, using sample size of 500 with 1000 replicates with coefficients of variation

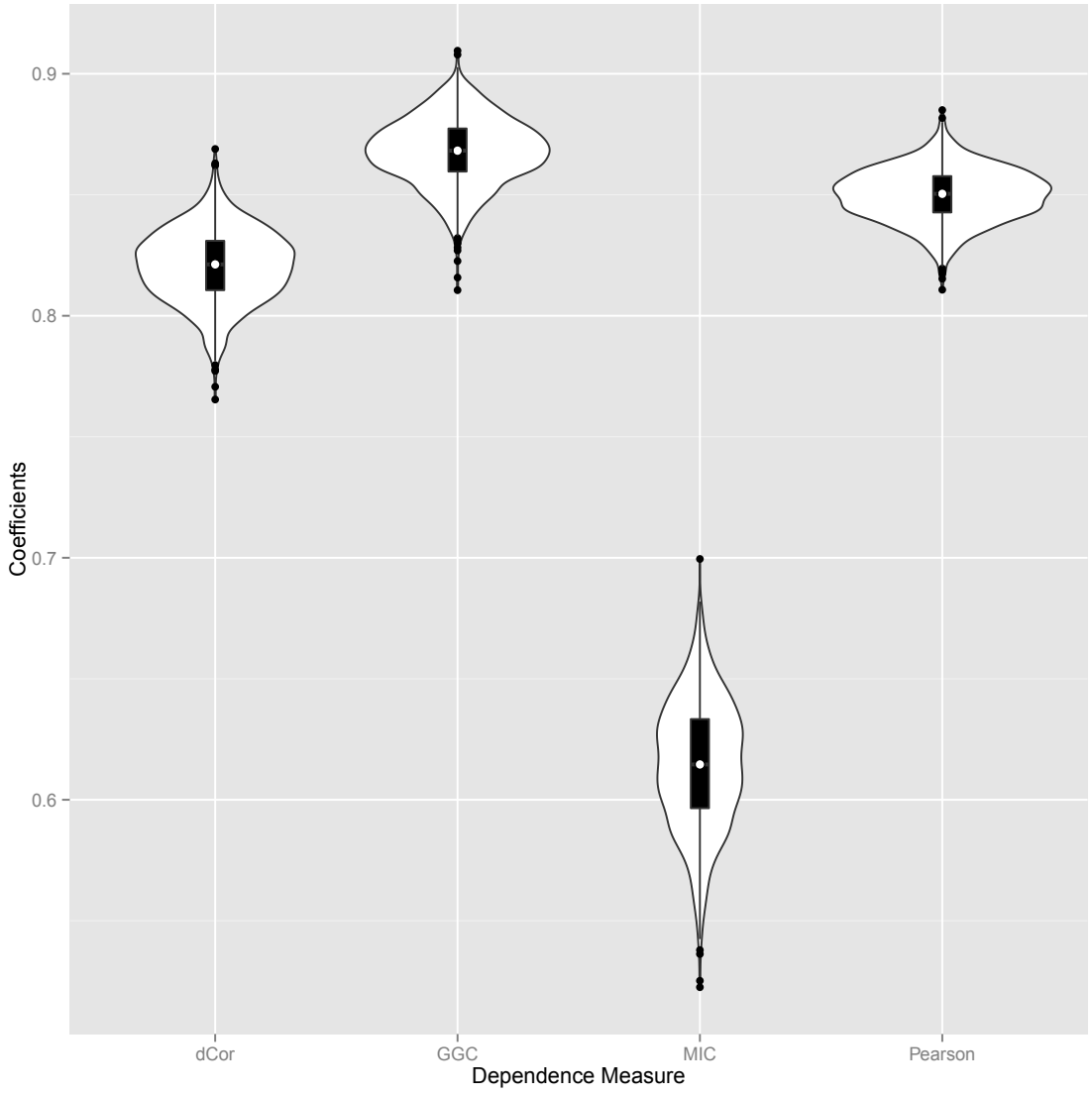
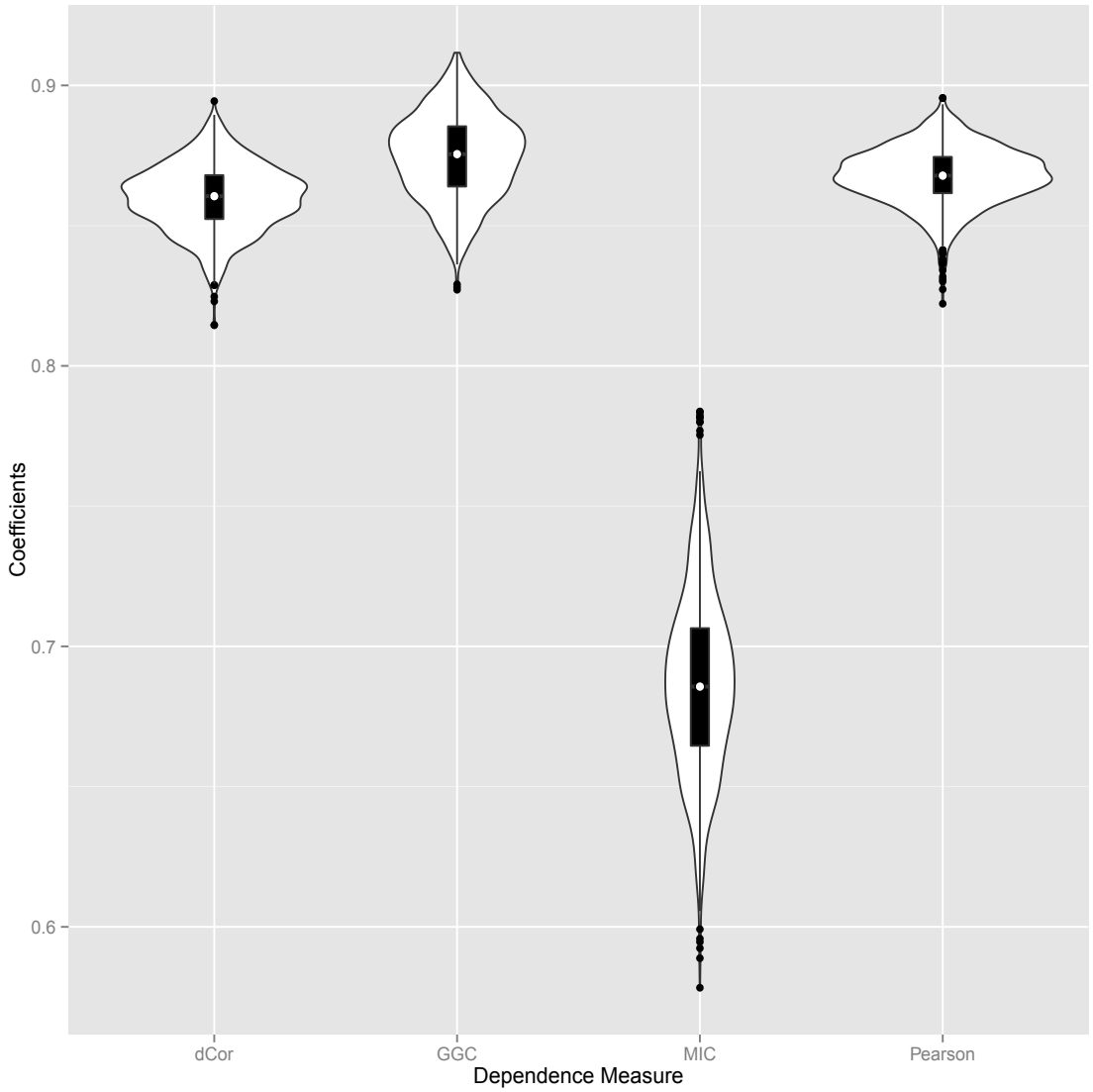


Figure 5.5: Comparison of the measures describing exponential dependence: $Y = 4\exp(0.5X) + 2 + \varepsilon$, where $X, \varepsilon \sim N(0, 1)$ are independent, using sample size of 500 with 1000 replicates



monotone.

5.1.5 Sinusoidal model

For the sinusoidal data, the distribution of X is uniformly distributed at $(0, 1)$ and $Y_i = \sin(4\pi X_i) + \varepsilon_i$, where ε_i are standard uniformly distributed and independent of X. The model is plotted in Figure 5.1e. The violin plot in Figure 5.6 shows that MIC estimates perform well in detecting this type of relationship, giving an average coefficient of about 0.87. The coefficients of dCor and GGC are equally close on average to 0.4, but dCor has smaller variance. Pearson's r, on the other hand, detects a negative linear relationship, which is inaccurate.

5.1.6 Diamond Data

The diamond data example was taken from Newton [66], whose R-code can be found in Appendix A. An illustration of the diamond plot can be seen in Figure 5.1f. The violin plot presented in Figure 5.7 illustrates that MIC and dCor capture a weak association between X and Y, with their means centered on 0.17 and 0.14, respectively. GGC suggests no association with mean at 0.095. The sampling distribution of Pearson's r clusters and centers around 0, which means that most of the time it cannot detect any relationship. This is consistent with its characteristic of identifying linearity only.

5.1.7 Four Independent Clouds

The four independent clouds example was also taken from Newton [66], whose R-code can be found in Appendix A. An illustration of the four independent clouds can be seen in Figure 5.1g. The violin plot displayed in Figure 5.8 illustrates that the sampling distributions of dCor and r cluster around 0, with dCor estimates being closer to 0 than Pearson's r. Thus, both dCor and r are consistent with independence between X and Y for this type of structure. However, the sampling distributions of GGC and MIC statistics reveal some type of relationship which centers at 0.21 and 0.16, respectively. With the nature of GGC, being a sum of ρ^2, it may be capturing some dependence of points within a neighborhood or reflecting the fact that ρ varies a great deal over

Figure 5.6: Comparison of the measures describing sinusoidal dependence: $Y = \sin(4\pi X) + \varepsilon$, where $X, \varepsilon \sim U(0, 1)$ using sample size of 500 with 1000 replicates

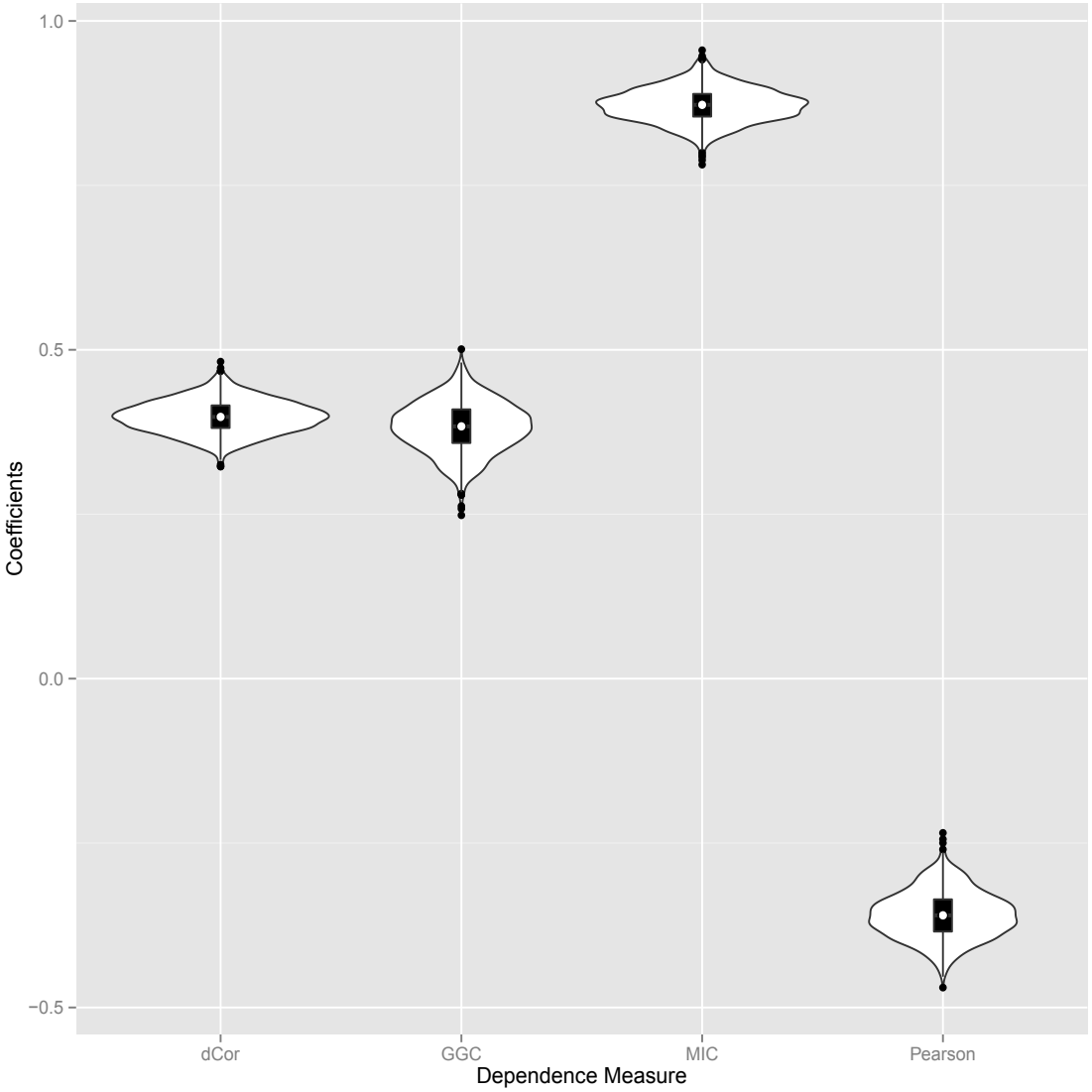


Figure 5.7: Comparison of the measures describing diamond relationship using Newton's [66] R-code with a sample size of 500 and 1000 replicates

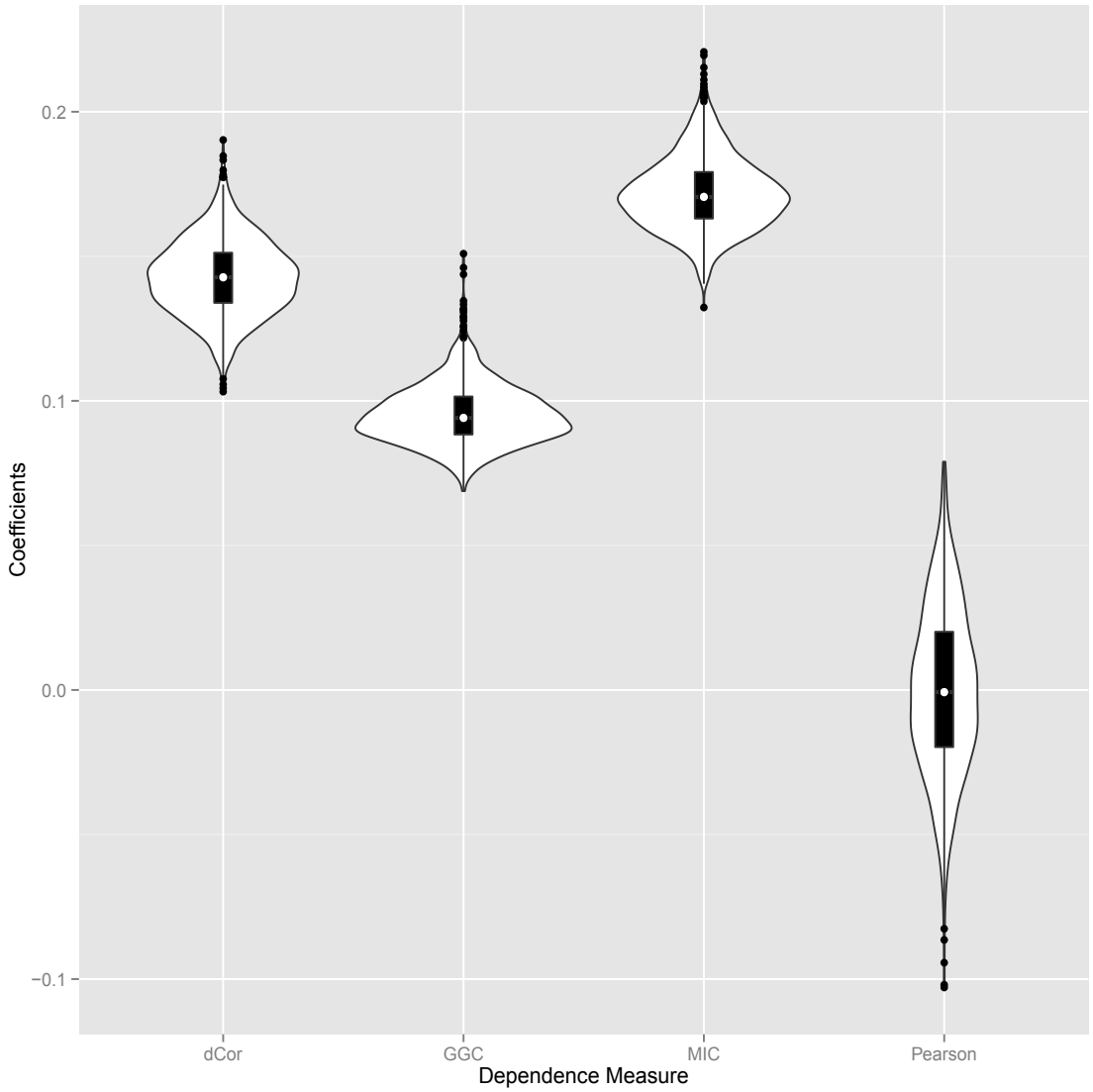
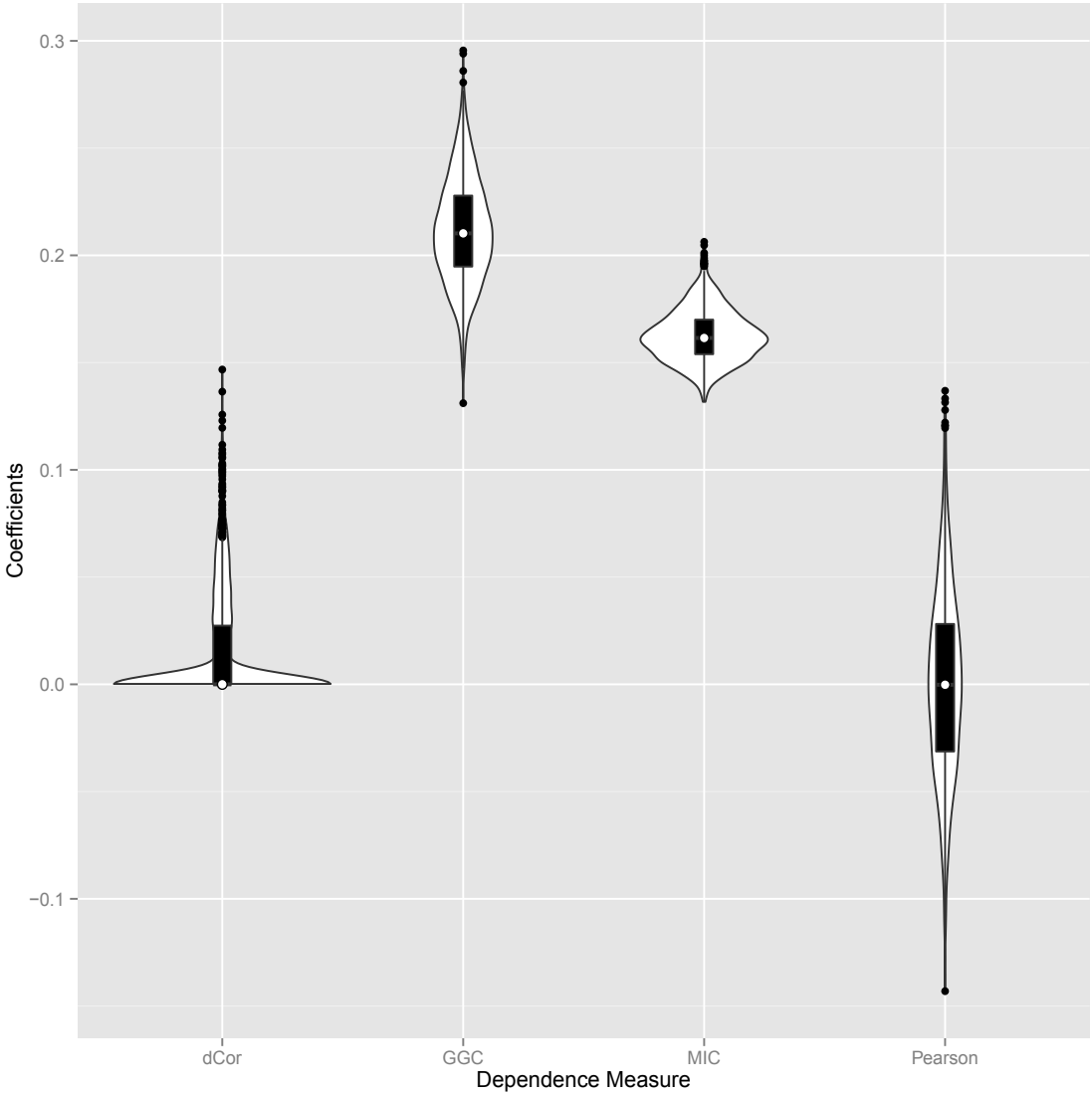


Figure 5.8: Comparison of the measures describing four independent clouds using Newton's [66] R-code with a sample size of 500 and 1000 replicates



the whole grid on $\mathbb{R}^2$, even though X and Y are independent. The estimates of GGC are highly dispersed than the estimates of MIC.

5.1.8 Independent bivariate t model

In this model, the random variables X and Y are independent and identically t-distributed with $v = 4$ degrees of freedom. Figure 5.1h exhibits the plot. The violin plot displayed in Figure 5.9 illustrates the sampling distributions of each of the four methods, with the summary of the estimates found in Table 5.1. It is easily observed that both dCor and Pearson perform better in quantifying independence of the two random variables since the distributions are both centered around 0. Centers of MIC and GGC are far from 0. Furthermore, the coefficients of GGC are highly variable, ranging from 0.2 to 0.6, which suggests that GGC may overstate the dependence when the marginal distributions have heavy tails. This is caused by the way GGC is constructed. Local Gaussian correlation estimates dependence in a neighborhood of each point x, and if the points are close to each other, it assigns dependence which contributes to the global measure. With heavy tailed data, the density estimates have greater error in the tails, which can skew the global estimate.

5.2 Statistical Power and Type-I Error Rates

In this section, we summarize the power comparisons of the dependence measures for bivariate and multivariate independence.

While dCor, HHG and RV tests are valid in arbitrary dimension, Pearson's r is only applicable for bivariate data, and GGC has only been implemented for bivariate data. Thus, we consider bivariate association in Section 5.2.1. Multivariate association is investigated in Section 5.2.2 for dCor, HHG and RV coefficients.

5.2.1 Bivariate Association

The same dependence structures in Section 5.1 are considered and those measures which are applicable for bivariate associations were compared.

Figure 5.9: Comparison of the measures describing t-independent random variables with $v = 4$ degrees of freedom using sample size of 500 with 1000 replicates

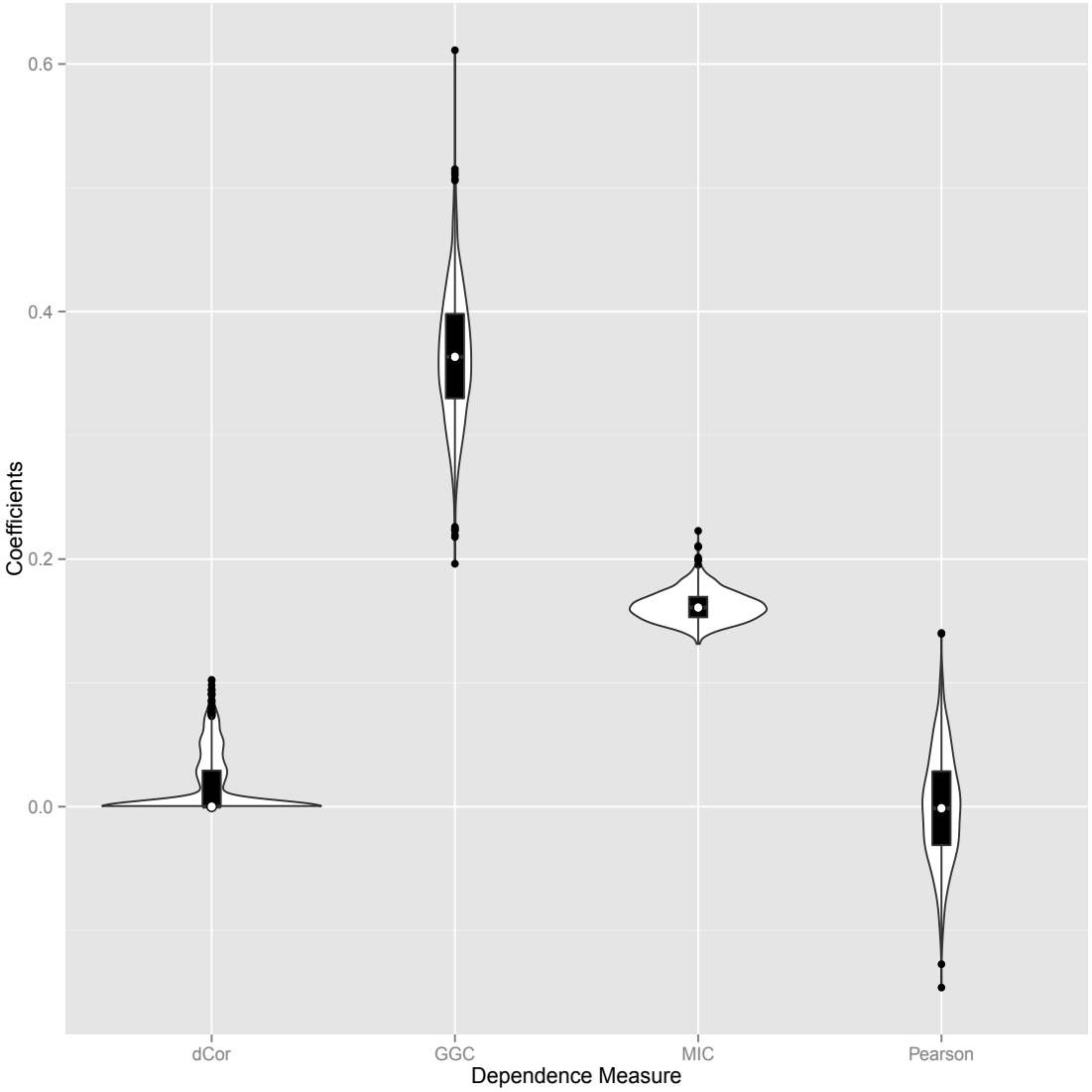


Table 5.2: Empirical Type-I error rates (with standard error in parentheses) for 1000 tests at nominal significance level 0.05 of two independent structures

Independence Structure	Sample Size	dCor	GGC	HHG	Pearson
Four Independent Clouds	25	0.042 (0.006)	0.050 (0.007)	0.043 (0.006)	0.049 (0.007)
	35	0.040 (0.006)	0.052 (0.007)	0.045 (0.007)	0.046 (0.007)
	50	0.048 (0.007)	0.052 (0.007)	0.057 (0.007)	0.050 (0.007)
	65	0.047 (0.007)	0.053 (0.007)	0.049 (0.007)	0.044 (0.006)
	75	0.041 (0.006)	0.043 (0.006)	0.045 (0.007)	0.041 (0.006)
	100	0.046 (0.007)	0.045 (0.007)	0.042 (0.006)	0.050 (0.007)
Independent t	25	0.062 (0.008)	0.057 (0.007)	0.056 (0.007)	0.059 (0.007)
	35	0.054 (0.007)	0.054 (0.007)	0.049 (0.007)	0.060 (0.008)
	50	0.049 (0.007)	0.059 (0.007)	0.040 (0.006)	0.052 (0.007)
	65	0.053 (0.007)	0.055 (0.007)	0.051 (0.007)	0.039 (0.006)
	75	0.049 (0.007)	0.062 (0.008)	0.043 (0.006)	0.043 (0.006)
	100	0.052 (0.007)	0.050 (0.007)	0.045 (0.007)	0.044 (0.006)

Table 5.2 shows the Type-I error rates of two examples of a null distribution: the four independent clouds, and the two independent random variables X and Y which are both t-distributed. Figures 5.10 and 5.11 display the plot. It can be seen that all four measures correctly control the Type-I error rate at the nominal significance level $\alpha = 0.05$.

Table 5.3 displays the power performance of the bivariate measures. As shown, GGC is not as powerful as dCor, HHG statistics and Pearson's r when the relationship is linear. The low power of GGC may be due to the fact that reliable density estimates of the local Gaussian distribution require moderately large sample size in each neighborhood. The four measures have approximately the same empirical power, which is very strong in detecting a cubic relationship. They are also as powerful when capturing sinusoidal dependence. Similar results are obtained for the exponential model except when the sample size n is less than 50. In detecting quadratic dependence, dCor and HHG are equally powerful while GGC and Pearson's r are less powerful. This is due to the fact that r is designed to capture linearity only. An unusual relationship like the diamond can be detected by HHG test.

5.2.2 Multivariate Association

There are three multivariate measures that we compare. These are distance correlation (dCor), HHG test and the RV coefficient. There are only a few multivariate measures that are consistent against all dependent alternatives, and these include dCor and HHG.

Table 5.4 displays the empirical Type I error rates of random vectors X and Y which are distributed under the null for 1000 tests at varying sample sizes. First, the marginal distributions of X and Y are standard multivariate normal in dimensions $p = q = 5$. Next, the random vectors X and Y are generated from the $t(v)$ distribution when $v = 2, 3$. All three multivariate measures correctly controlled Type I error rates at the nominal significance level $\alpha = 0.05$.

Following Székely et al. [91], we consider two examples of nonlinear dependence between two five-dimensional random vectors ($p = 5$). In Figure 5.12, the distribution of X is standard multivariate normal, and $Y_{ij} = log(X_{ij}^2)$, where $j = 1, \ldots, p$. In Figure 5.13, the distribution of X is standard multivariate normal, and $Y_{ij} = X_{ij}\varepsilon_{ij}$, where $j = 1, \ldots, p$. As depicted in the figures,

Table 5.3: Power(with standard error in parentheses) of the four dependence measures using various sample sizes in different dependence structures. Results are based on 1000 simulations

Dependence Structure	Sample Size	dCor	GGC	HHG	Pearson
Linear	25	1.000 (0)	0.092 (0.009)	0.999 (0)	1.000 (0)
	35	1.000 (0)	0.141 (0.011)	1.000 (0)	1.000 (0)
	50	1.000 (0)	0.254 (0.014)	1.000 (0)	1.000 (0)
	65	1.000 (0)	0.370 (0.015)	1.000 (0)	1.000 (0)
	75	1.000 (0)	0.450 (0.016)	1.000 (0)	1.000 (0)
	100	1.000 (0)	0.654 (0.015)	1.000 (0)	1.000 (0)
Quadratic	25	0.978 (0.005)	0.015 (0.004)	0.998 (0.001)	0.366 (0.015)
	35	1.000 (0)	0.021 (0.005)	1.000 (0)	0.376 (0.015)
	50	1.000 (0)	0.047 (0.007)	1.000 (0)	0.366 (0.015)
	65	1.000 (0)	0.081 (0.009)	1.000 (0)	0.382 (0.015)
	75	1.000 (0)	0.105 (0.010)	1.000 (0)	0.360 (0.015)
	100	1.000 (0)	0.182 (0.012)	1.000 (0)	0.363 (0.015)
Cubic	25	1.000 (0)	0.983 (0.004)	1.000 (0)	1.000 (0)
	35	1.000 (0)	0.999 (0.001)	1.000 (0)	1.000 (0)
	50	1.000 (0)	0.999 (0.001)	1.000 (0)	1.000 (0)
	65	1.000 (0)	1.000 (0)	1.000 (0)	1.000 (0)
	75	1.000 (0)	1.000 (0)	1.000 (0)	1.000 (0)
	100	1.000 (0)	1.000 (0)	1.000 (0)	1.000 (0)
Exponential	25	1.000 (0)	0.594 (0.016)	1.000 (0)	1.000 (0)
	35	1.000 (0)	0.796 (0.013)	1.000 (0)	1.000 (0)
	50	1.000 (0)	0.916 (0.009)	1.000 (0)	1.000 (0)
	65	1.000 (0)	0.972 (0.004)	1.000 (0)	1.000 (0)
	75	1.000 (0)	0.991 (0.003)	1.000 (0)	1.000 (0)
	100	1.000 (0)	1.000 (0)	1.000 (0)	1.000 (0)
Sinusoid	25	0.583 (0.016)	0.488 (0.016)	0.794 (0.013)	0.446 (0.016)
	35	0.810 (0.012)	0.611 (0.015)	0.981 (0.004)	0.581 (0.016)
	50	0.984 (0.004)	0.780 (0.013)	1.000 (0)	0.771 (0.013)
	65	0.999 (0.001)	0.882 (0.010)	1.000 (0)	0.871 (0.011)
	75	1.000 (0)	0.912 (0.009)	1.000 (0)	0.908 (0.009)
	100	1.000 (0)	0.975 (0.005)	1.000 (0)	0.914 (0.009)
Diamond	25	0.030 (0.005)	0.008 (0.003)	0.211 (0.013)	0.006 (0.002)
	35	0.034 (0.006)	0.011 (0.003)	0.366 (0.015)	0.004 (0.002)
	50	0.047 (0.007)	0.011 (0.003)	0.656 (0.015)	0.004 (0.002)
	65	0.063 (0.008)	0.014 (0.004)	0.846 (0.011)	0.003 (0.002)
	75	0.068 (0.008)	0.008 (0.003)	0.918 (0.009)	0.002 (0.001)
	100	0.135 (0.011)	0.015 (0.004)	0.981 (0.004)	0.002 (0.001)

Table 5.4: Empirical Type-I error rates for 10000 tests at nominal significance level 0.05 for three multivariate examples involving two independent multivariate normal X and Y and two multivariate t-distributed X and Y with degrees of freedom $v = 2, 3$

Structure	Sample Size	dCor	HHG	RV
Multivariate Normal	30	0.0538 (0.0023)	0.0491 (0.0022)	0.0532 (0.0022)
$p = q = 5$	40	0.0493 (0.0022)	0.0532 (0.0022)	0.0504 (0.0022)
	50	0.0501 (0.0022)	0.0471 (0.0021)	0.0513 (0.0022)
	60	0.0505 (0.0022)	0.0491 (0.0022)	0.0501 (0.0022)
	70	0.0516 (0.0022)	0.0499 (0.0022)	0.0519 (0.0022)
	80	0.0510 (0.0022)	0.0531 (0.0022)	0.0514 (0.0022)
	90	0.0505 (0.0022)	0.0484 (0.0021)	0.0487 (0.0022)
	100	0.0507 (0.0022	0.0499 (0.0022)	0.0508 (0.0022)
t(v=2)	30	0.0513 (0.0022)	0.0488 (0.0022)	0.0549 (0.0023)
	40	0.0519 (0.0022)	0.0498 (0.0022)	0.0544 (0.0023)
	50	0.0485 (0.0021)	0.0514 (0.0022)	0.0463 (0.0021)
	60	0.0502 (0.0022)	0.0542 (0.0023)	0.0491 (0.0022)
	70	0.0483 (0.0021)	0.0488 (0.0022)	0.0435 (0.0020)
	80	0.0494 (0.0022)	0.0498 (0.0022)	0.0409 (0.0020)
	90	0.0504 (0.0022)	0.0521 (0.0022)	0.0457 (0.0021)
	100	0.0524 (0.0022)	0.0529 (0.0022)	0.0411 (0.0020)
t(v=3)	30	0.0610 (0.0024)	0.0474 (0.0021)	0.0521 (0.0022)
	40	0.0624 (0.0024)	0.0501 (0.0022)	0.0495 (0.0022)
	50	0.0622 (0.0024)	0.0516 (0.0022)	0.0515 (0.0022)
	60	0.0567 (0.0023)	0.0430 (0.0020)	0.0455 (0.0021)
	70	0.0598 (0.0024)	0.0513 (0.0022)	0.0465 (0.0021)
	80	0.0602 (0.0024)	0.0479 (0.0021)	0.0440 (0.0021)
	90	0.0588 (0.0024)	0.0473 (0.0021)	0.0446 (0.0021)
	100	0.0597 (0.0024)	0.0496 (0.0022)	0.0455 (0.0021)

Figure 5.10: Empirical Type-I error rates of dCor, GGC, HHG, and Pearson for 1000 tests at nominal significance level $\alpha = 0.05$ for four independent clouds.

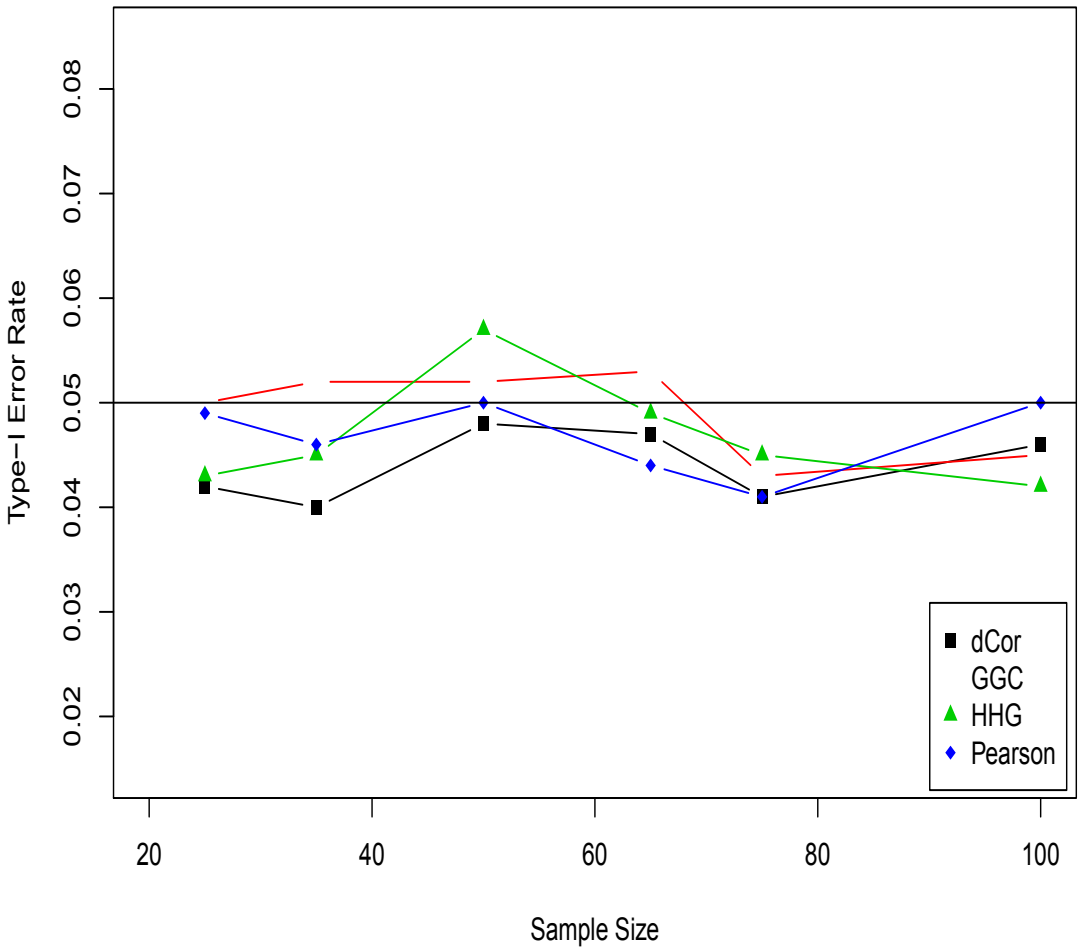


Figure 5.11: Empirical Type-I error rates of dCor, GGC, HHG, and Pearson for 1000 tests at nominal significance level $\alpha = 0.05$ for two independent t-distributed samples.

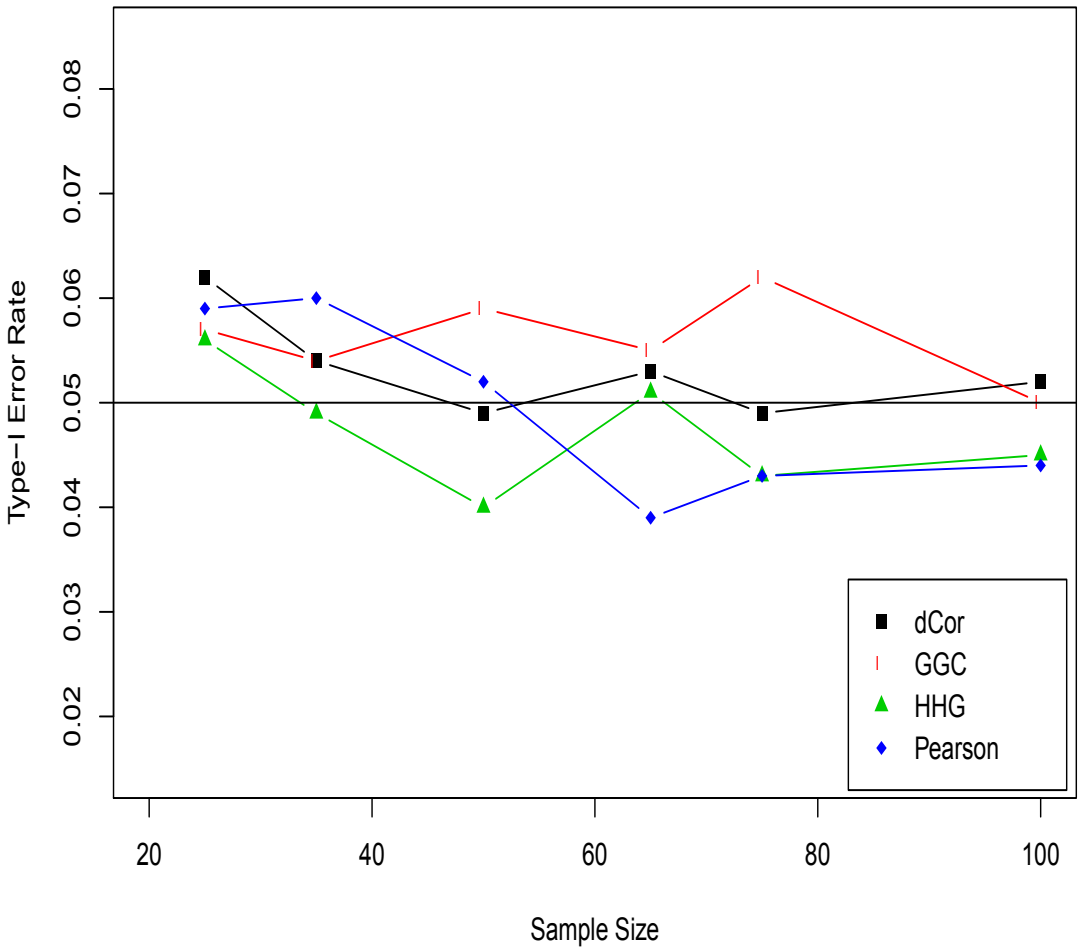
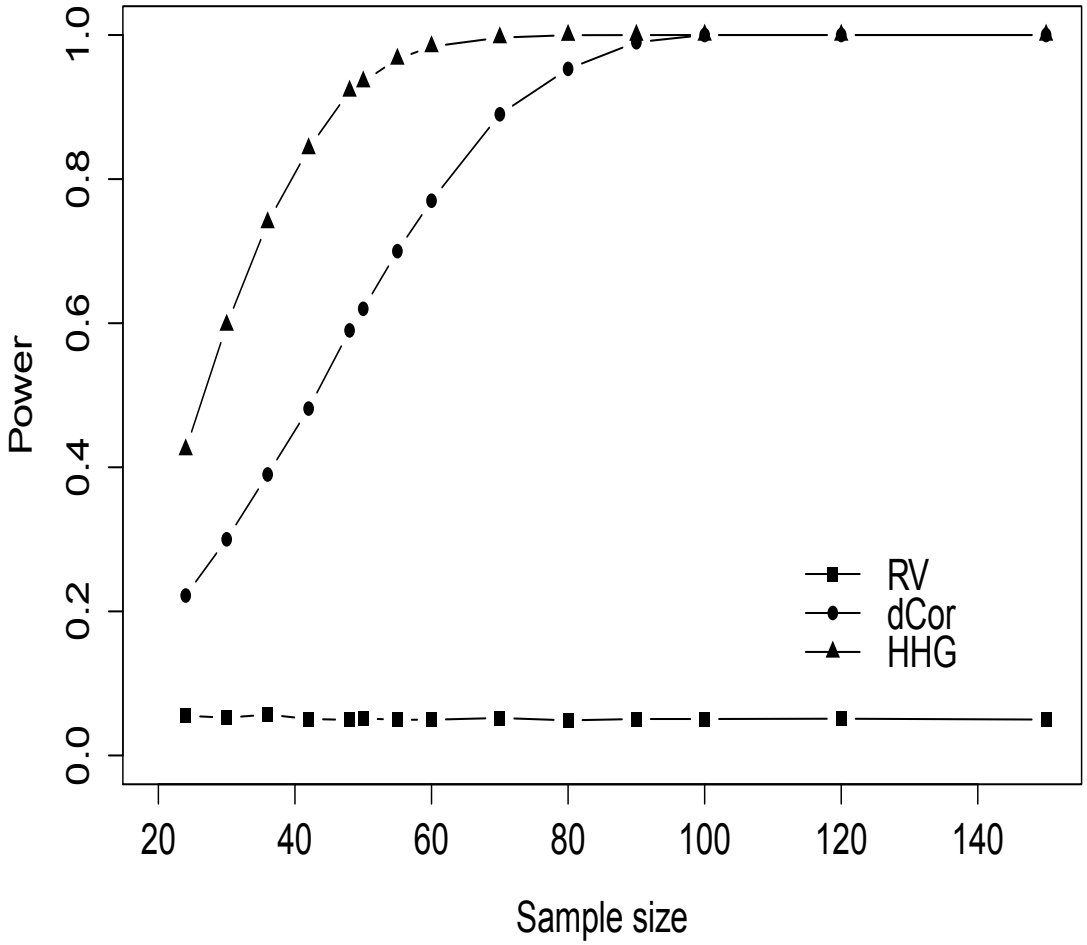
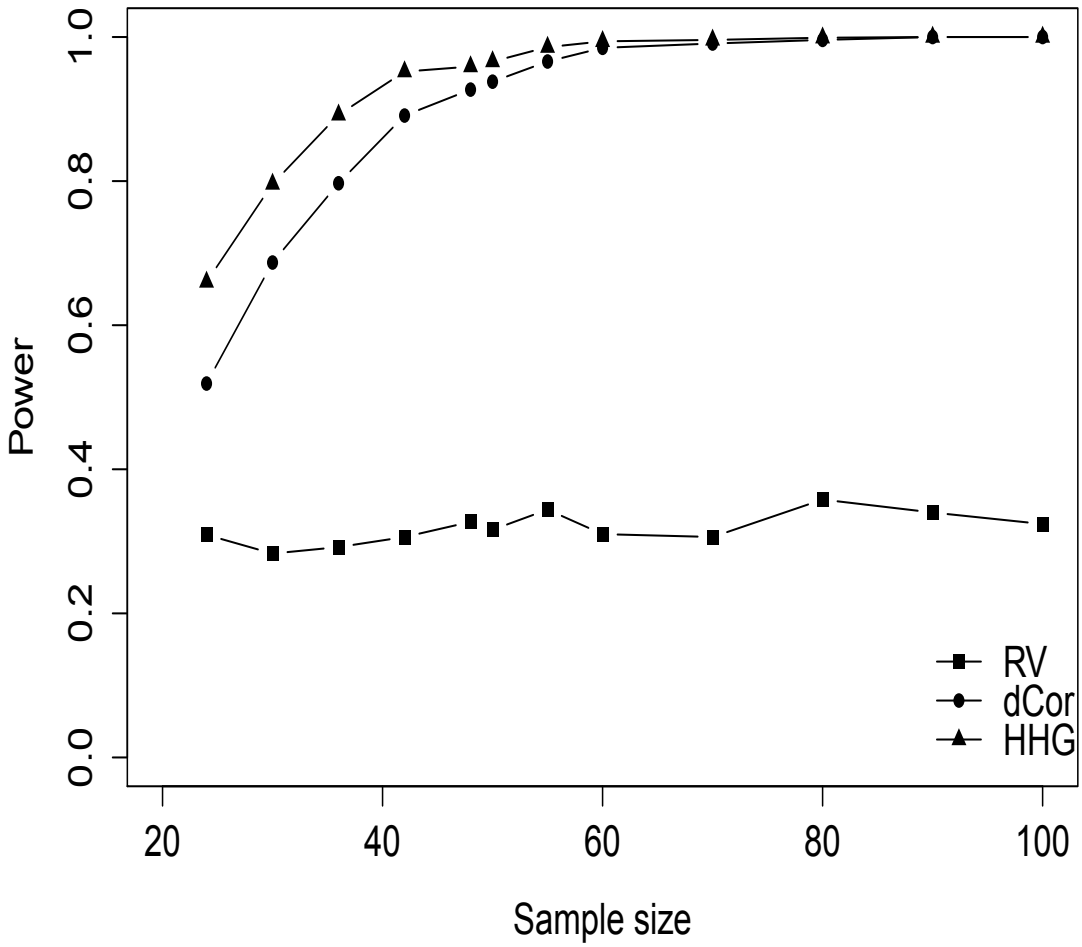


Figure 5.12: Empirical power for dCor, HHG, the RV tests at nominal significance level $\alpha = 0.05$ when $Y = log(X^2)$. Results are based on 10000 simulations



dCor and HHG are indeed comparably powerful in both scenarios, but RV has lower power. Recall that the RV is designed to measure linear association.

Figure 5.13: Empirical power for dCor, HHG, RV tests at nominal significance level $\alpha = 0.05$ when $Y = XE$. Results are based on 10000 simulations



CHAPTER 6 APPLICATION

6.1 Bivariate Example on Total Stock Returns

Most people study many different variables involved in the stock market before they invest their money. In fact, they want to know which stocks are more profitable. Knowledgeable people look at the performance of total stock returns and dividends, and perhaps other related financial data. However, relationships of these variables change over time and season. Hence, it is best to figure out which variables are essentially associated to each other to come up with the best decision. It is necessary to use a measure that can accurately detect possible dependence.

In this section, we explore the dependence structure of a real multivariate financial data set that is pertinent to the stock market. The data set is taken from Yahoo!Finance [102] on March 12, 2014. A total of 399 companies from the 500 largest companies having common stock listed on the New York Stock Exchange (NYSE) or National Association of Securities Dealers Automated Quotations (NASDAQ), had a complete data for the variables considered below. These 399 companies were examined. Table 6.1 gives the financial variables with their definitions taken from [35] that we considered for each stock in the S&P 500 Index.

In particular, we want to know which of the above variables are related to total stock return. The formula for the total stock return as defined by Hirschey [43] is the appreciation in the price plus any dividends paid, divided by the original price of the stock. The income sources from a stock are dividends and its increase in value. The first portion of the numerator of the total stock return formula looks at how much the value has increased from the initial price P_0 to the closing price P_1. The denominator of the formula to calculate a stock's total return is the original price of the stock, which is the original amount invested. The annual total returns of a stock (TSR) is given by

$$TSR = \frac{(P_1 - P_0) + D}{P_0},$$

where P_0 is the initial stock price, P_1 is the ending stock price (after 1 year), and D is the annual dividends.

The scatterplot matrix in Figure 6.1 gives a picture of the association between each pair of variables for the 399 stocks with a loess line smoother, a nonparametric regression method. Of more interest is to determine which financial variable is related to total stock returns (TSR) during the period March 13, 2014 to March 12, 2015. To analyze the stock return data, each stock was considered as a bivariate observation with each of the variables in Table 6.1 as the X component and TSR as the Y component. Then, each financial variable X was paired to the TSR and the three test statistics were compared. The statistics of dCov as given in Equation (3.1.11) with their corresponding p-values are computed using the function `dcov.test` with 999 replicates found in the *energy* package in R. The coefficients of LGauss given in Equation (3.2.5) are computed using the function `global.lgauss`. To obtain the corresponding p-values, 499 random permutations are simulated using the function `global.lgausstest` which can be found in Appendix A. The MIC statistics are calculated using the function `mine` from the *minerva* package [24] in R while the pre-computed p-values of the MIC scores when $n = 399$ are taken from the MINE website in [74] under *Downloads*. The Pearson r coefficients with their corresponding p-values are determined using function `cor.test`. The results of the statistics and the corresponding p-values are shown in Table 6.2. A significance level of 0.05 is used in this example.

As seen in Table 6.2, the tests for dCov, GGC, MIC and Pearson's r give the same conclusion about its relationship with total stock returns (TSR) when the independent variables are *Percent Annual Dividend Yield (ADY)*, *Earnings per Share (EPS)* and *Payout Ratio (POR)*. All measures except Pearson product moment correlation reveal that there exists a significant association of *Market Capital (MC)* with the total stock returns. All except MIC has detected relationship with *Quarterly Revenue Growth (QRG)*. Both GGC and MIC uncover association of stock returns with *Enterprise Value (EV)* but dCor and Pearson did not. Only dCor and GGC detect relationship with *Price/Earnings to Growth Ratio (PEG)* and *Revenues per Share (RPS)*. However, it is worth noting that only dCor detects *Price-to-Sales (PS)*, *Payout Ratio (POR)* and *Percent Held by Institutions*

Table 6.1: Financial variables considered for evaluation with their definitions

Financial Variable	Definition
Market Capital (MC)	A measure of a company's total value; the company's share price multiplied by the number of shares a company has outstanding.
Enterprise Value (EV)	A more comprehensive alternative to equity market capitalization, which includes debt, minority interest, and preferred shares.
Price/Sales (PS)	A valuation ratio that compares a company's stock price to its revenues. The price-to-sales ratio is an indicator of the value placed on each dollar of a company's sales or revenues.
Percent Annual Dividend Yield (ADY)	A financial ratio that shows how much a company pays out in dividends each year relative to its share price. In the absence of any capital gains, the dividend yield is the return on investment for a stock.
Earnings per Share (EPS)	The amount of net sales a company achieves per share of common stock issued and outstanding.
Price/Earnings to Growth Ratio (PEG)	A stock's price-to-earnings ratio divided by the growth rate of its earnings for a specified time period. Used to determine a stock's value while taking the company's earning growth into account.
Revenues per Share (RPS)	The amount of net sales a company achieves per share of stock issued and outstanding.
Quarterly Revenue Growth (QRG)	An increase of a company's sales when compared to a previous quarter's revenue performance.
Payout Ratio (POR)	The percentage of a company's earnings paid out to shareholders in the form of dividends.
Percent Held by Institutions (PBI)	Percentage of outstanding common shares being held by institutional investors.

Figure 6.1: Scatterplot matrix of pairwise associations of some of the variables used in fundamental analysis of stocks of the S&P 500 index for the period 2014-2015.

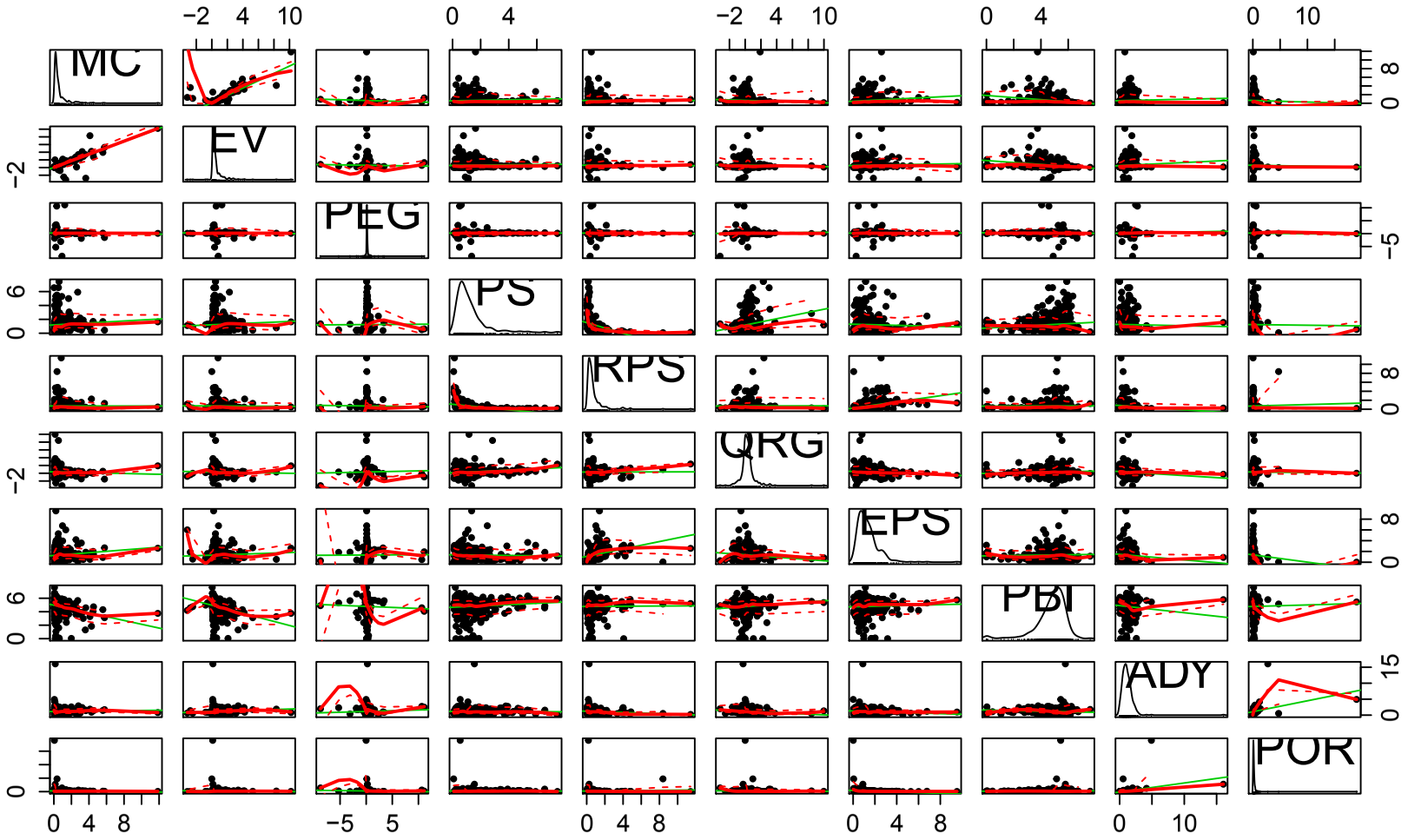


Table 6.2: Computed test statistics (with p-values in parentheses) of the four dependence measures testing bivariate relationship of each financial variable with Total Stock Return

Financial Variable	dCov	LGauss	MIC	Pearson
MC	1.5297 (0.0370)	0.3621 (0.0392)	0.2036 (0.0460)	0.0813 (0.1049)
EV	1.1118 (0.1370)	0.4364 (0.0195)	0.2290 (0.0018)	0.0471 (0.3485)
PS	2.9919 (0.0100)	0.3686 (0.0784)	0.1803 (0.0600)	0.0854 (0.0883)
ADY	2.9097 (0.0020)	0.3549 (0.0196)	0.2263 (0.0026)	-0.1297 (0.0095)
EPS	0.5334 (0.9820)	0.2654 (0.6078)	0.1871 (0.0600)	0.0085 (0.8654)
PEG	0.9437 (0.0160)	0.6695 (0.0196)	0.1939 (0.0600)	0.0239 (0.6334)
RPS	2.4639 (0.0010)	0.3855 (0.0392)	0.1908 (0.0600)	0.0572 (0.2545)
QRG	6.1778 (0.0010)	0.3448 (0.0193)	0.1958 0.0600)	0.2645 (8e-08)
POR	0.7548 (0.0160)	0.3141 (0.1569)	0.1873 (0.0600)	0.0597 (0.2342)
PBI	2.2620 (0.0340)	0.2719 (1.0000)	0.1942 (0.0600)	0.0464 (0.3552)

(PBI) as significantly correlated to total stock returns.

This specific example about stock returns reflects that different dependence measures can behave differently. They perform based on the nature of their construction, but some measures can really capture significant relationships with high power just like dCor.

6.2 Multivariate Example on Valuation Measures and Stock Trading Information

In relation to the financial statistics in the previous bivariate example , we test whether there is an association between valuation measures and trading information of stocks. Two components of valuation measures consisting of *Market Capital* and *Enterprise Value* are included as variables in X. Two components of trading information consisting of the *52-week Change* and the *200-day Moving Average* are included as variables in Y. There are 399 companies observed which makes X a 399×2 vector and Y a 399×2 vector.

We are looking at multivariate dependence between X and Y so the measures, that are applicable to assess the relationship, are dCor, HHG and RV. The dCov test statistic and its corresponding p-value is computed using the function `dcov.test` with 999 replicates in the *energy* package [77] in R. The HHG test statistic given in Equation (3.5.1) and the p-value using 1000 random permutations is calculated using the function `hhg.test` implemented in the *HHG* package [14] in R. The RV coefficient and the Pearson type III approximation to test its significance is determined using the function `coeffRV` in *FactoMineR* [47] package in R. The tests are analyzed using a significance level of 0.05. The results in Table 6.3 show that dCor and HHG detected a significant association between valuation measures and trading information of stocks but RV did not. This implies that trading information is affected by the company's shares and market values.

6.3 Multivariate Example on Parkinson's Disease

This example deals with telemonitoring Parkinson's disease as discussed by Tsanas, Little, McSharry, and Ramig [100].

According to Ronken and van Scharrenburg [80], Parkinson's disease (PD) tends to be regarded

Table 6.3: Computed statistics with p-values of the multivariate measures testing association of Valuation Measures vs Trading Information at 0.05 significance level

Dependence Measure	Test Statistic	p-value
dCov	$n\mathcal{V}_n^2 = 4.4213$	0.0040000
HHG	$\chi^2 = 300504.8$	0.0039801
RV	$RV = 0.00863203$	0.09764239

as a disease linked to aging. Ages between 50 and 60 years old are the most cases diagnosed, and less than 10% is identified below the age 40. There are 3.5M cases of PD world-wide that are recognised by World Health Organization (WHO), making it the most common cause of long-term disability in the elderly. PD, known as "shaking pals", is a degenerative disorder of the central nervous system. It is characterised by symptoms that are primarily affecting the motor system such progressively developing tremor, rigidity, slowness of movement and postural instability. Later, thinking and behavioral problems may arise, with dementia commonly occurring in the advanced stages of the disease, whereas depression is the most common psychiatric symptom. Other symptoms include sensory, sleep and emotional problems. Tsanas et al. [100] stated that the progression of PD is currently monitored by the Unified Parkinson's Disease Rating Scale (UPDRS) which includes analysis of speech of the patient. The clinician assesses whether the subject's vocal output is understandable and/or expressive during casual conversation by looking at the jitter and shimmer of the patient. According to Farrús and Hernando [25], jitter and shimmer are measures of the fundamental frequency and amplitude cycle-to-cycle variations, respectively. Both features have been largely used for the description of pathological voices, and since they characterize some aspects concerning particular voices, they are expected to have a certain degree of speaker specificity.

The data set is taken from UCI Machine Learning Repository [61]. There are 5875 patients, in which five measures of variation in fundamental frequency and six measures of variation in amplitude were observed. We consider the X random vector with 5875 rows and 5 columns where the columns consist of the five measures of variation in fundamental frequency: Jitter, Jitter:Abs, Jit-

Table 6.4: Computed statistics with p-values of the multivariate measures testing association of variation in fundamental frequency vs variation in amplitude at 0.05 significance level

Dependence Measure	Test Statistic	p-value
dCov	$n\mathcal{V}_n^2 = 1.1400000$	0.001000
HHG	$\chi^2 = 14012650262$	0.000997
RV	$RV = 0.4916228$	0.000000

ter:RAP, Jitter:PPQ5, Jitter:DDP. We consider the Y random vector with 5875 rows and 6 columns where the columns consist of the six measures of variation in amplitude: Shimmer, Shimmer:dB, Shimmer:APQ3, Shimmer:APQ5, Shimmer:APQ11, Shimmer:DDA. We test, at 0.05 level of significance, whether there is an association between the measures of variation in fundamental frequency $X_{5875\times5}$ and the measures of variation in amplitude $Y_{5875\times6}$.

The measures that are used to assess for multivariate dependence between X and Y are dCor, HHG and RV. The dCov test statistic and its corresponding p-value is computed using the function `dcov.test` with 999 replicates in the *energy* package [77] in R. The HHG test statistic given in Equation (3.5.1) and the p-value using 1000 random permutations is calculated using the function `hhg.test` implemented in the *HHG* package [14] in R. The RV coefficient and the Pearson type III approximation to test its significance is determined using the function `coeffRV` in *FactoMineR* package [47] in R. The tests are analyzed using a significance level of 0.05.

As shown in Figure 6.4, the dCov test of independence of X and Y, HHG test statistic and RV coefficient all detect a significant multivariate relationship. Results may indicate a linear component in the dependence structure because a dependence was detected by RV, a generalization of the Pearson product-moment correlation. This implies that there is a significant association between the measures of variation in fundamental frequency and the measures of variation in amplitude.

CHAPTER 7 SUMMARY

We have compared some of the modern bivariate and multivariate measures of dependence that are currently well-known in the statistical community. These measures are distance correlation, global Gaussian correlation, maximal information coefficient, RV coefficient and HHG test.

The distance covariance (dCov) and distance correlation (dCor) can be utilized to describe bivariate and multivariate associations. They are similar to the covariance and correlation developed by Pearson, but more general, since they can detect both linear and nonlinear dependence of any random vectors X and Y, for any distribution in arbitrary dimension as long as first moments of X and Y are finite. A generalization, according to Székely and Rizzo [94], makes dCov and dCor applicable for any X and Y with finite α-moments, for some $\alpha > 0$.

The global Gaussian correlation (GGC or τ), which aggregates the local Gaussian correlation on a subset of $\mathbb{R}^2$, recognizes linear and nonlinear dependence structures in bivariate data. The global coefficient is more useful when the data are normally distributed. The local Gaussian correlation can specifically distinguish negative and positive dependence for bivariate data. Currently, it is implemented only for bivariate data.

Maximal information coefficient (MIC) is an exploratory data analysis tool developed for identifying interesting relationships of several pairs of variables and characterizing these relationships according to properties such as nonlinearity and monotonicity.

The RV coefficient is a multivariate generalization of the squared Pearson correlation coefficient, which measures linear relationship. Similar to the Pearson product-moment correlation, $\rho V = 0$ does not necessarily imply that X and Y are independent unless an assumption of multivariate normal is satisfied.

The HHG statistic determines a nonparametric test used to detect associations between random vectors of any dimension. It makes use of ranks of distances and is a consistent test against all dependent alternatives.

The empirical results show how the different measures behave and treat different dependence

structures. All of the measures studied perform well in detecting linear, cubic and exponential association, but MIC ranks last among them for these types of structures. All measures except Pearson's r are able to capture quadratic and diamond relationship, with the global Gaussian correlation performing best. All measures except Pearson's r, with higher coefficients of MIC, are able to detect sinusoidal association. Distance correlation and Pearson's r have correct Type I error rates in the examples of Section 5.1, but the other measures do not always have results consistent with independence.

It was shown that the dCov test is as powerful in our simulations as HHG test in detecting most of the bivariate relationships examined, including linear, quadratic, cubic, exponential and sinusoid models. Both are more powerful in our simulations than GGC and Pearson's r in revealing quadratic and sinusoidal relationships. All are equally powerful in our simulations in identifying the existence of a pure linear association.

For detecting multivariate association, dCor and HHG are equally powerful in our simulations. Both are consistent against all dependence alternatives and the tests achieve good power for finite sample sizes. Not many multivariate tests of independence available at present have these properties.

Three dependence measures, namely, dCor, GGC and RV, are assessed according to the properties of Rényi including Pearson's ρ and $|\rho|$. None of the measures satisfied all seven properties of Rényi but there are some properties that are partially fulfilled. Only two to three properties of Rényi are fully satisfied by the measures.

Many other desirable properties are examined such as applicability in high-dimensional, scale invariance, consistency. Most specifically, properties such as rigid motion invariance and equitability are discussed. Rigid motion invariance is a property that an interpretable dependence measure should hold. It is important that even if you translate, rotate or reflect the data points, without changing any lengths or angles between the points, the value of the measure should not change. Only dCor and RV fulfill this property. Kinney and Atwal [54] investigated that no dependence measure has possessed the equitability property that Reshef, Reshef, Finucane, Grossman,

McVean, Turnbaugh, Lander, Mitzenmacher and Sabeti proposed. In addition, Kinney and Atwal proposed the possibility of dependence measures having the properties of self-equitability and Data Processing Inequality, however, none of the measures satisfy them as well.

As in the findings of Kinney and Atwal [54], we have shown by example that no measure satisfied the property of equitability. Gorfine et al. [36] mentioned that the property of equitability does not help when a test has low power and cannot detect much. We conclude that equitability is not an essential property for a dependence measure.

The results we have observed in this study are consistent with the statements provided by some authors and experts on these tests and measures. In a comment by Gorfine et al. [36], MIC is relevant only for bivariate data while HHG and dCor work also in a multivariate setting. They have found that the dCor and HHG tests are more powerful than the test based on MIC, and thus are preferable over MIC. In addition to this, in our results, dCor is more powerful than GGC and Pearson's r in identifying several bivariate dependencies including monotonic and nonmonotonic relationships.

In comparing measures and tests of independence, both theoretically and empirically, we have seen that dCor and GGC meet most of the properties that a dependence measure should possess, although the other measures and tests have their own advantages. However, GGC is suitable for bivariate data only while dCor is good for bivariate and multivariate data with any distribution and any dimension. Distance correlation is as powerful or sometimes more powerful in detecting association than competing bivariate and multivariate tests popular today. Simon and Tibshirani [85] believe that dCor is a more powerful technique that is simple, easy to compute, and should be considered for general use.

For further studies, we would recommend investigating, comparing and evaluating the partial correlations, if there are any, of these recent measures of dependence. If possible, one can also develop a partial correlation for any of them which doesn't have yet. In addition, we suggest to establish a dependence measure specific to time series data as there are not many dependence measure that is available for this data type. Another feasible research is to propose a measure

of correlation for high dimensional data. Lastly, we would recommend to explore more of the copula-based dependence measures and compare their advantages and disadvantages with non copula-based.

BIBLIOGRAPHY

[1] Abdi, H. (2007). RV coefficient and congruence coefficient. In (Eds.) N.J. Salkind, editor. *Encyclopedia of Measurement and Statistics*. Sage, Thousand Oaks, CA. 849-853.

[2] Atwood, J. and Spolsky, J. (2014). Correlation coefficient distribution function. Stack Exchange Inc. URL *http://math.stackexchange.com/questions/681226/correlation-coefficient-distribution-function-an-apparent-discrepancy*.

[3] Bakirov, N. K., Rizzo, M. L. and Székely, G. J. (2006). A multivariate nonparametric test of independence. *Journal of Multivariate Analysis* (97) 1742-1756.

[4] Balakrishnan, N., Lai, C. (2009). Continuous bivariate distributions. Springer, New York. 152.

[5] Berentsen, G., Kleppe, T., Tjøstheim, D. (2014). Introducing localgauss, an R package for estimating and visualizing local Gaussian correlation. *Journal of Statistical Software* **56(12)** 1-18.

[6] Berentsen, G. and Tjøstheim, D. (2014). Recognizing and visualizing departures from independence in bivariate data using local Gaussian correlation. *Statistics and Computing* **24(5)** 785-801.

[7] Bhuchongkul, S. (1964). A class of nonparametric tests for independence in bivariate populations. *Annals of Mathematical Statistics*. **35** 138-149.

[8] Bilodeau, M. and Lafaye de Micheaux, P. (2005). A multivariate empirical characteristic function test of independence with normal marginals. *Journal of Multivariate Analysis* **95** 345-369.

[9] Bilodeau, M. and Lafaye de Micheaux, P. (2012). Nonparametric tests of independence between random vectors. URL *http://cran.fhcrc.org/web/packages/IndependenceTests/IndependenceTests.pdf*.

[10] Blomqvist, N. (1950). On a measure of dependence between two random variables. *Annals of Mathematical Statistics* **21** 593-600.

[11] Blum, J., Kiefer, J., Rosenblatt, M. (1961). Distribution free tests of independence based on the sample distribution function. *Annals of Mathematical Statistics* **32** 485-498.

[12] Bobko, P. (2001). Correlation and regression: applications for industrial organizational psychology and management. *SAGE Publications* ISBN 150631595X, 9781506315959 12-26.

[13] Breiman, L., Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation: rejoinder. *Journal of the American Statistical Association* **80** 614-619.

[14] Brill, B., Kaufman, S. (2015). Heller-Heller-Gorfine tests of independence and equality of distributions. URL *http://cran.r-project.org/web/packages/HHG/HHG.pdf*.

[15] Cléroux, R., Ducharme, G.R.(1989). Vector correlation for elliptical distribution. *Communications in Statistics Theory and Methods* **18** 1441-1454.

[16] Cléroux, R., Lazraq, A., Lepage, Y.(1995). Vector correlation based on ranks and a nonparametric test of no association between vectors. *Communications in Statistics Theory and Methods* **24** 713-733.

[17] Cover, T., Thomas, J. (1991) Elements of Information Theory. John Wiley and Sons, New York.

[18] Csorgo, S. (1985). Testing for independence by empirical characeristic function. *Journal of Multivariate Analysis* **16** 290-299.

[19] Delicado, P. Smrekar, M. (2008). Measuring nonlinear dependence for two random variables distributed along a curve. *Journal of Multivariate Analysis* **16** 290-299.

[20] Dray, S. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*. **22(4)** 1-20.

[21] Escoufier, Y. (1970). Echantillonnage dans une population de variables aleatoires reelles. *Publ. Inst. Stat. Univ. Paris* **19** (fasc. 4) 1-47.

[22] Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics* **29** 751-760.

[23] Feuerverger, A. (1993). A consistent test for bivariate dependence. *International Statistical Review* **61** 419-433.

[24] Filosi, M., Visintainer, R., Albanese, D., Riccadonna, S., Jurman, G., Furlanello, C. (2014). minerva: Maximal Information Based Nonparametric Exploration R. URL *http://cran.r-project.org/package=minerva*, R package version 1.4.1.

[25] Farrus, M. and Hernando, J. (2009). Using jitter and shimmer in speaker verification. *IET Signal Process*. **3** 247-257.

[26] Fisher, R. (1954). *Statistical Methods for Research Workers*. Oliver and Boyd.

[27] Friendly, M. (2002). Corrgrams: Exploratory data analysis for correlation matrices. URL *http://euclid.psych.yorku.ca/datavis/papers/corrgram.pdf*.

[28] Galton, F. (1888). Co relations and their measurement, chiefly from anthropometric data. *Proceedings of the Royal Society of London*. **45** 135-145.

[29] Galton, F. (1889). *Natural Inheritance*. Macmillan, Chapters 4 and 5.

[30] Gebelein, H. (1941) Das statistiche problem der korrelation als variantions and eigenwert-problem und sein zusammennhang mit der ausgleichsrechnung. *Z. Angew. Math. Mech.* **21** 364-379.

[31] Geiser, P. and Randles, R. (1997) A nonparametric test of independence between two vectors. *Journal of American Statistical Association* **92** 561-567.

[32] Gelman, A. (2012). Statistical modeling, causal inference and social science. URL *http://andrewgelman.com/2012/03/26/further-thoughts-on-nonparametric-correlation-measures/*.

[33] Gelman, A. (2013). Statistical modeling, causal inference and social science. URL *http://andrewgelman.com/2013/02/04/what-principles-should-govern-attempts-to-summarize-bivariate-associations-in-large-multivariate-datasets/*.

[34] Gelman, A. (2014). Statistical modeling, causal inference and social science. URL *http://andrewgelman.com/2014/03/14/maximal-information-coefficient/*.

[35] Gildersleeve, R. (1999). Winning business: how to use financial analysis and benchmarks to outscore your competition. Cashman Dudley. ISBN 9780884158981.

[36] Gorfine, M., Heller, R., Heller, Y. (2012) Comment on"Detecting Novel Associations in Large Data Set". URL *http://www.tau.ac.il/ gorfinem/pdf/Gorfine2012a.pdf*.

[37] Gower, J. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*. **53** 325-338.

[38] Granger, C., Maasoumi, E., Racine,J. (2004). A dependence metric for possibly nonlinear processes. *Journal of Time Series Analysis*. **25** 649-669.

[39] Gretton, A., Herbrich, R., Smola, A., Bousquet, O., Scholkopf, B. (2005). Kernel methods for measuring dependence. *Journal of Machine Learning Research* **6** 2075-2129.

[40] Gretton, A., Bousquet, O., Smola, A., Scholköpf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. *Lecture notes in computer science*. Algorithmic Learning Theory **3734** 63-77.

[41] Hall, W. J. (1970). On characterizing dependence in joint distributions. In: Essays in Probability and Statistics. R. C. Bose, I. M. Chakravarti, P. C. Mahalanobis, C. R. Rao, and K. J. C. Smith, editors. University of North Carolina Press, Chapel Hill. 339-376.

[42] Heller, R., Heller, Y., Gorfine, M. (2012). A consistent multivariate test of association based on ranks of distances. *Biometrika* 1-8.

[43] Hirschey, M. (2003). Tech Stock Valuation: Investor Psychology and Economic Analysis. Academic Press. London.

[44] Hjort, N. Jones, M. (1996). Locally parametric nonparametric density estimation. *Annals of Statistics* **24(4)** 1619-1647.

[45] Hoeffding, W. (1948). A nonparamteric test of independence. *The Annals of Mathematical Statistics* **19** 546-557.

[46] Hotelling, H. and Pabst, M. (1936). Rank correlation and tests of significance involving no assumption of normality. *The Annals of Mathematical Statistics* **7** 29-43.

[47] Husson, F. Josse, J., Le, S. Mazet,J. (2013). FactoMineR: Multivariate Exploratory Data Analysis and Data Mining with R. URL http://cran.r-project.org/package=FactoMineR, R package version 1.24.

[48] Hutter, M. and Zaffalon, M. (2005). Distribution of mutual information from complete and incomplete data. *Computational Statistics and Data Analysis* **48** 633-657.

[49] Jones, M., Koch, I. (2003). Dependence maps: local dependence in practice. *Statistics and Computing* **13** No. 3 241.

[50] Josse, J., Pages, J. Husson, F. (2008). Testing the significance of rv coefficient. *Computational Statistics and Data Analysis* **53** 82-91.

[51] Josse, J., Holmes, S. (2013). Measures of dependence between random vectors and tests of independence. URL *http://arxiv-web.arxiv.org/pdf/1307.7383v2.pdf*.

[52] Kazi-Aoual, F., Hitier, S., Sabatier, R., Lebreton, J.D. (1995). Refined approximations to permutation tests for multivariate inference. *Computational Statistics and Data Analysis* **20** 643-656.

[53] Kendall, M. (1938). A new measurement of rank correlation. *Biometrika* **30** 81-93.

[54] Kinney, J., Atwal, G. (2014). Equitability, mutual information and the maximal correlation coefficient. *PNAS* **111** 3354-3359.

[55] Kraskov, A., Stogbauer, H., Grassberger, P. (2004). Estimating mutual information. *Physical Review E Statistical, Nonlinear, and Soft-Matter Physics* **69** 1-15.

[56] Kruskal, W. (1958). Ordinal measures of dependence. *Journal of American Statistical Association* **53** 814-861.

[57] Lancaster, H. O.(1963). Correlation and complete dependence of random variables. *Annals of Mathematical Statistics* **34** 1315-1321.

[58] Le S., Josse, J., Husson, F. (2008). FactoMineR: an R package for multivariate analysis. *Journal of Statistical Software*. **25(1)** 1-18.

[59] Lehmann, E. (1966). Some concepts of dependence. *Annals of Mathematical Statistics* **37** 1137-1153.

[60] Lehmann, E. L. (2011). Fisher, Neyman, and the Creation of Classical Statistics. Springer Science and Business Media, Bucher, Germany.

[61] Lichman, M. (2013). UCI Machine Learning Repository. URL *http://archive.ics.uci.edu/ml*.

[62] Liebetrau, A. M. (1983). Measures of Association. SAGE Publications, New York.

[63] Linfoot E. H. (1957). An informational measure of correlation. *Information and Control*. **1** 85-89.

[64] Lyons, R. (2013). Distance covariance in metric spaces. *Annals of Probability*. **41** 3284-3305.

[65] Maddala, G. S. (2001). Introduction to Econometrics. Wiley, New York. ISBN 0471497282, 9780471497288. 155.

[66] Newton, M. A. (2009). Introducing the discussion paper by Szekely and Rizzo. *The Annals of Applied Statistics*. **3** 1233-1235.

[67] Pearson, K. (1896). Mathematical contributions to the theory of evolution III. Regression, Heredity and Panmixia. *Philosophical Transactions of the Royal Scoiety A*. **187**. 253-318.

[68] Puri, M. and Sen, P. (1971) Nonparametric Methods in Multivariate Analysis. Wiley, New York.

[69] R Development Core Team. (2014). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL *http://www.r-project.org/*.

[70] Rahman, N. A. (1968). A Course in Theoretical Statistics. Charles Griffin and Company, London.

[71] Reimherr, M., Nicolae, D. (2013). On quantifying dependence: a framework for developing interpretable measures. *Statistical Science* **28** 160-130.

[72] Rényi, A. (1959). On measures of dependence. *Acta. Math. Acad. Sci. Hungary* **10** 441-451.

[73] Rényi, A. (1970). Probability Theory. Charles Griffin and Company. North-Holland, Amsterdam.

[74] Reshef D., Reshef Y. (2011). Maximal Information based Nonparametric Exploration (MINE). URL *http://www.exploredata.net*.

[75] Reshef, D., Reshef, Y., Finucane, H., Grossman, S., McVean, G., Turnbaugh, P., Lander, E., Mitzenmacher, M., Sabeti, P.C. (2011). Detecting novel associations in large data sets. *Science* **334** 1518-1524.

[76] Reshef D., Reshef Y., Mitzenmacher M., Sabeti P. (2013). Equitability analysis of the maximal information coefficient with comparisons. URL *http://arxiv.org/pdf/1301.6314v2.pdf*.

[77] Rizzo, M.L., Székely, G.J. (2013). E statistics (energy statistics). URL *http://cran.r-project.org/package=energy*

[78] Robert, P., Cleroux, R., Ranger, N. (1985). Some results on vector correlation. *Computational Statistics and Data Analysis* **3** 724-732.

[79] Robert, P., Escoufier, Y. (1976). A unifying tool for linear multivariate statistical methods: The RV-coefficient. *Journal of the Royal Stat. Society Series C (Applied Statistics)* **3** 257-265.

[80] Ronken, E., Van Scharrenburg, G.J.M. (2002). Parkinson's disease. *IOS Press)*. Amsterdam, The Netherlands.

[81] Schweizer, B., Wolff, E. (1981). On nonparametric measures of dependence for random variables. *The Annals of Statistics* **9** No.4 879-885.

[82] Sethuraman, J. (1990). The asymptotic distribution of Rényi maximal correlation. *Comm. Statist. Theory Methods* **19** 4291-4298.

[83] Shannon, C., Weaver, W. (1949). The mathematical theory of communication. University of Illinois Press, Urbana, Illinois.

[84] Siburg, K., Stoimenov, P. (2010). A measure of mutual complete dependence. *Metrika* **71** No.2 239-251.

[85] Simon, N., Tibshirani, R. (2011) Comment on "Detecting Novel Associations in Large Data Sets" by Reshef et al. (Science Dec 2011). URL *http://statweb.stanford.edu/ tibs/reshef/comment.pdf*.

[86] Sinha, B., Wieand, H. (1977). Multivariate nonparametric tests for independence. *Journal of Multivariate Analysis*. **7(4)** 572-583.

[87] Sklar, A. (1973). Random variables, joint distribution functions and copulas. *Kybernitika* **9** 449-460.

[88] Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology* **15** 72-101.

[89] Spector, P., Friedman, J., Tibshirani, R., Lumley, T. (2014). acepack: ace() and avas() for choosing regression transformations. URL *http://CRAN.R-project.org/package=acepack*.

[90] Stove, B., Tjøstheim, D., Hufthammer, K. (2014). Using local Gaussian correlation in a nonlinear re-examination of financial contagion. *Journal of Empirical Finance* **25** 62-82.

[91] Székely, G.J., Rizzo, M.L. and Bakirov, N.K. (2007). Measuring and testing independence by correlation of distances. *Annals of Statistics* **35** 2769-2794.

[92] Székely, G.J., Rizzo, M.L. (2009). Brownian distance covariance. *The Annals of Applied Statistics* **3(4)** 1236-1265.

[93] Székely, G.J., Rizzo, M.L. (2012). On the uniqueness of distance covariance. *Statistics and Probability Letters* **82** 2278-2282.

[94] Székely, G.J., Rizzo, M.L. (2013). The distance correlation t-test of independence in high dimension. *Journal of Multivariate Analysis* **117** 193-213.

[95] Székely, G.J., Rizzo, M.L. (2013). Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference* **143** 1249-1272.

[96] Székely, G.J., Rizzo, M.L. (2014). Partial distance correlation with methods for dissimilarities. *The Annals of Statistics* **42** 2382-2412.

[97] Taskinen, Kakainen, A., Oja, H. (2003) Sign test of independence between two random vectors. *Statistics and Probabilty Letters*. **62** 9-21.

[98] Taskinen, S., Oja, H. Randles, R. (2005). Multivariate nonparametric tests of independence. *Journal of the American Statistical Association* **100** 916-925.

[99] Tjøstheim, D., Hufthammer, K. (2013). Local Gaussian correlation: a new measure of dependence. *Journal of Econometrics* **172(1)** 33-48.

[100] Tsanas, A., Little, M., McSharry, P. and Ramig, L. (2009). Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. *IEEE transactions on biomedical engineering*. **57** No. 4 884-893.

[101] Wilks, S.S. (1935). On the independence of k sets of normally distributed statistical variables. *Econometrica* **3** 309-326.

[102] Yahoo contributors. (2015). Yahoo finance. URL *http://finance.yahoo.com/q/ks?s=ABC+key+Statistics*.

[103] Yenigün, D. (2007). A test of independence in two-way contingency tables based on maximal correlation. PhD Thesis *Bowling Green State University*.

APPENDIX A SELECTED R PROGRAMS

R-codes: A. The following are used to simulate different dependence structures such linear, parabolic, cubic, exponential, sinusoid, and independent *t*.

```
simlin<-replicate(M, expr={
        X <- rnorm(500,sd=2.5)
        e <- rnorm(500,sd=2.5)
        Y <- 2*X+e
        list=c(X=X,Y=Y)
        })
simqua<-replicate(1000, expr={
        X <- rnorm(500,sd=1.5)
        e <- rnorm(500,sd=1.5)
        Y <- X^2+e
        list=c(X=X,Y=Y)
        })
simcub<-replicate(1000, expr={
       X <- runif(500,-1.3,1)
       e <- rnorm(500,mean=1.5,sd=0.85)
       Y <- 4*X^3+ X^2+e
       list=c(X=X,Y=Y)
       })
simlp<-replicate(1000, expr={
       X <- runif(500)
       e <- runif(500)
       Y <- sin(4*pi*X)+e
```

```
        list=c(X=X,Y=Y)
        })
simexp<-replicate(1000, expr={
        X <- rnorm(500)
        e <- rnorm(500)
        Y <- 4*exp(0.5*X)+2+e
        list=c(X=X,Y=Y)
        })
simt<-replicate(1000, expr={
        X <- rt(500,4)
        Y <- rt(500,4)
        list=c(X=X,Y=Y)
        })
```

B. The function *global.lgauss* is used to compute the global Gaussian correlation statistic Tn of a bivariate random variable X and Y. It makes use of the localgauss function that can be found in the package localgauss. The function *global.lgausstest* is a permutation test of independence based on the global Gaussian correlation that outputs the p-value.

```
global.lgauss<-function(x,y){
  lg.out<-localgauss(x,y)
  tn<-sqrt(sum(lg.out$par.est[,5]^2)/NROW(lg.out$par.est))
  return(list(tn=tn))
}
global.lgausstest<-function(x,y,n,R){
  tn0<-global.lgauss(x,y)$tn
  z<-c(x,y)
  N<-2*n
  K<-1:N
```

```
  reps<-numeric(R)
  for(i in 1:R){
    k=sample(K, size=n, replace=FALSE)
    x1=z[k]
    y1=z[-k]
    reps[i]<-global.lgauss(x1,y1)$tn
  }
  pval<-(sum(reps>=tn0)+1)/(R+1)
  return(list(Tn=tn0, p.value=pval))
}
```

C. Example of Self-equitability and DPI of Global Gaussian correlation (LGauss):

```
set.seed(32477)
n<-100
epsilon<-runif(n,-0.5,0.5)
X<-runif(n,-2,2)
fx<-X^2
Y<-exp(X^2)+0.5+epsilon
global.lgauss(X,Y)$tn
global.lgauss(fx,Y)$tn
```

D. Example of global Gaussian correlation not being a rigid motion invariant:

```
set.seed(1234)
n<-50
t<-4
X1<-rt(n,t)
Y1<-rt(n,t)
global.lgauss(X1,Y1)
```

```
global.lgauss(3*X1,-4*Y1)
global.lgauss(5*X1+5,-4*Y1-1)
X2<-runif(n,-1,1)
epsilon<-rnorm(n,mean=2.5,sd=0.85)
Y2<-4*X2^3+X2^2+epsilon
global.lgauss(X2,Y2)
global.lgauss(3*X2,-4*Y2)
global.lgauss(5*X2+5,-4*Y2-1)
```

E. Example of MIC not being a rigid motion invariant:

```
set.seed(1234)
n<-50
t<-4
X1<-rt(n,t)
Y1<-rt(n,t)
mine(X1,Y1)$MIC
mine(-Y1,X1)$MIC
mine(-X1,-Y1)$MIC
mine(-X1+5,-Y1+3)$MIC
X2<-runif(n,-1.3,1)
epsilon2<-rnorm(n,mean=2.5,sd=0.85)
Y2<-4*X2^3+X2^2+epsilon2
mine(X2,Y2)$MIC
mine(-X2,-Y2)$MIC
mine(-3*X2,-4*Y2)$MIC
mine(-X2-5,-Y2+3)$MIC

set.seed(77744)
```

```
epsilon3<-runif(n,-1,1)
Y3<-sin(2*pi*X2)+X2+epsilon3
mine(X2,Y3)$MIC
mine(-X2,-Y3)$MIC
mine(-3*X2,-4*Y3)$MIC
mine(-X2-5,-Y3+3)$MIC
```

F. Multivariate Error rates:

```
library(energy)
library(FactoMineR)
library(HHG)
library(MASS)
mu <- c(0,0,0,0,0)
Sigma <- matrix(c(1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,0,1), 5, 5)
rmvn.eigen <-
    function(n, mu, Sigma) {
        d <- length(mu)
        ev <- eigen(Sigma, symmetric = TRUE)
        lambda <- ev$values
        V <- ev$vectors
        R <- V%*%diag(sqrt(lambda)) %*% t(V)
        Z <- matrix(rnorm(n*d), nrow = n, ncol = d)
        X <- Z%*%R + matrix(mu, n, d, byrow = TRUE)
        X
    }
errorates<-matrix(0,24,4)
size = c(seq(24,50,2), seq(55,100,5))
m<-1000
```

```
d<-5
alpha<-0.05
for(j in 1:length(size)){
      n<-size[j]
      pvalues<-replicate(m, expr={
                X <- rmvn.eigen(n, mu, Sigma)
                Y <- rmvn.eigen(n, mu, Sigma)
                rv<-coeffRV(X,Y)
                Dx<- as.matrix(dist(X),diag=TRUE,upper=TRUE)
                Dy<- as.matrix(dist(Y),diag=TRUE,upper=TRUE)
                dc<-dcov.test(X,Y,R=999)
                hhg<-hhg.test(Dx,Dy,nr.perm = 1000)
                c(dc$p.value,rv$p.value,hhg$perm.pval.hhg.sc)
                })
      errorates[j,1]<-n
      errorates[j,2]<-mean(pvalues[1,]<=alpha)
      errorates[j,3]<-mean(pvalues[2,]<=alpha)
      errorates[j,4]<-mean(pvalues[3,]<=alpha)
      }
errorates
```

G. M. A. Newton's R-code of diamond plot

```
x <- runif(n, min=(-1), max=1 )
y <- runif(n, min=(-1), max=1 )
theta <- -pi/4
rr <- rbind( c(cos(theta), -sin(theta) ),
             c( sin(theta), cos(theta) ) )
tmp <- cbind( x, y ) %*% rr
```

```
u <- tmp[,1]
v <-  tmp[,2]
xx[,i] <- u
yy[,i] <- v
```

H. M. A. Newton's R-code of four independent clouds plot

```
dx <- rnorm(n)/3
dy <- rnorm(n)/3
cx <- sample( c(-1,1), size=n, replace=T )
cy <- sample( c(-1,1), size=n, replace=T )
u <- cx + dx
v <- cy + dy
xx[,i] <- u
yy[,i] <- v
```